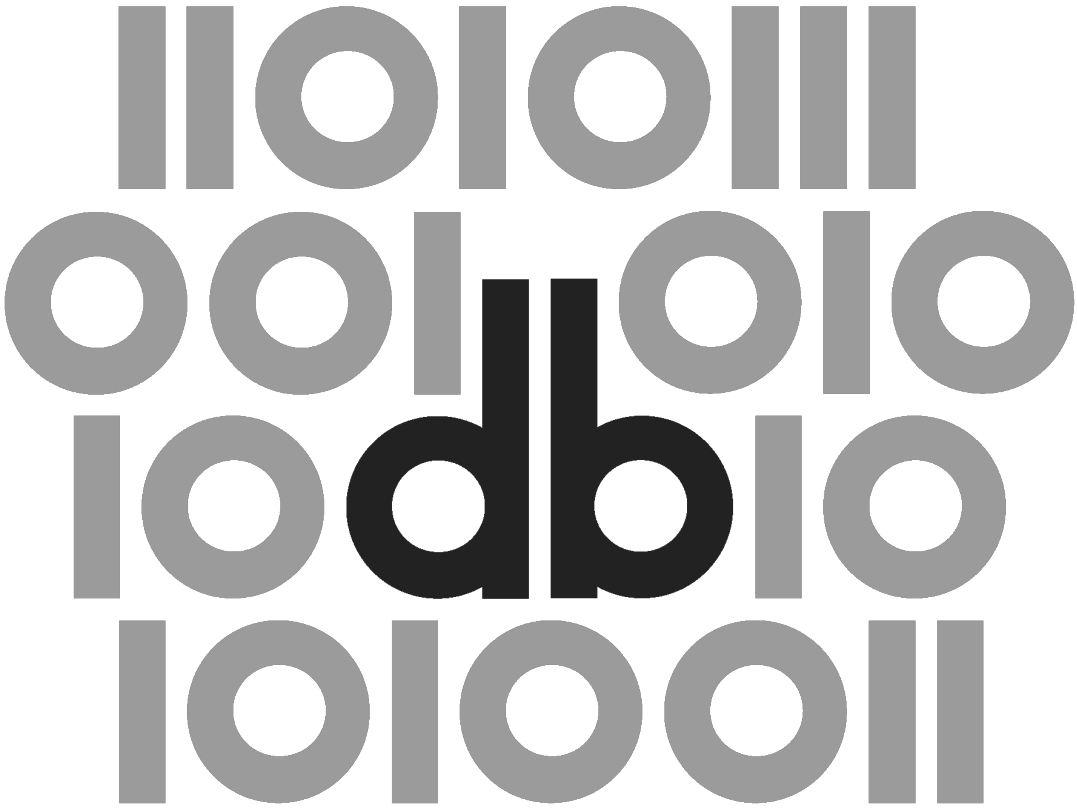
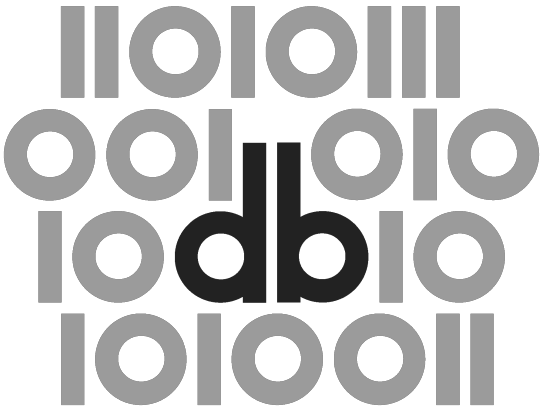
2 (2019) <DIGITÁLIS BÖLCSÉSZET>



2 (2019) </DIGITÁLIS BÖLCSÉSZET>

Digitális Bölcsészet
2019., második szám

<DIGITÁLIS BÖLCSÉSZET>



2 (2019)

ISSN 2630-9696
DOI 10.31400/dh-hun.2019.2
Kiadja az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A) és a Bakonyi Géza Alapítvány.
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Leváltár biztosítja.

Honlap: `http://ojs.elte.hu/index.php/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

Tartalom

<TANULMÁNYOK>

Koltay Tibor

Eszterházy Károly Egyetem

koltay.tibor@uni-eszterhazy.hu

Gondolatok a digitális bölcsészet, a könyvtártudomány és a könyvtárak kapcsolatrendszeréről*

A digitális bölcsészet területén dolgozó kutatók munkáját nem csak azért segíthetik a könyvtárak és a könyvtárosok, mert feladataik felmérésében és céljaik megvalósításában segítségükre van a szakértelmüket és tevékenységüket megalapozó könyvtár- és információtudomány (könyvtártudomány). Ez a két tudományterület a különbségek ellenére is számos hasonlóságot mutat egymással, mivel a könyvtártudomány célja, hogy elméleti szinten megismerje és megértse az információ világát és annak emberi tényezőit. Az utóbbi ugyanis valójában a digitális bölcsészeti kutatás tárgya is. Ennek a kérdéskörnek a tárgyalása mellett a cikk képet ad arról is, hogy – a nemzetközi szakirodalom tükrében – a világ (főként az Amerikai Egyesült Államok és Nagy-Britannia) egyetemi és szakkönyvtárai milyen szolgáltatásokat nyújtanak a digitális bölcsészeti kutatásokhoz.

Kulcsszavak:
könyvtár- és információtudomány, könyvtártudomány, egyetemi és szakkönyvtárak, könyvtári szolgáltatások, együttműködés

1. Bevezetés

Ahogy azt Sennyey Pongrácz kifejti, a digitális bölcsészet fogalma körüli viták újra és újra fellángolnak, és közben az érvek egyre kifinomultabbá és árnyaltabbá válnak.[1] A következőkben egy még kifinomultabb és árnyaltabb diskurzushoz szeretnék hozzájárulni azzal, hogy – főként a köztük levő hasonlóságok és különbségek vonatkozásában – betekintést nyújtok a digitális bölcsészet és a könyvtár- és információtudomány (röviden könyvtártudomány) kapcsolatrendszerébe. Emellett röviden bemutatom, hogy milyen képet fest a nemzetközi szakirodalom egy része a világ (főként az Amerikai Egyesült Államok és Nagy-Britannia) egyetemi és szakkönyvtárainak azon szolgáltatásairól, amelyeket a digitális bölcsészet területén dolgozó kutatóknak nyújtanak.

* A cikk megírását az EFOP-3.6.1-16-2016-00001 „Kutatási kapacitások és szolgáltatások komplex fejlesztése az Eszterházy Károly Egyetemen” projekt támogatta.

1 Sennyey Pongrácz, „Viták és víziók a digitális bölcsészetről,” *Digitális Bölcsészet* 1, 1. sz. (2018): 111–120, https://doi.org/10.31400/dh–hun.2018.1.228.

2. A digitális bölcsészet és a könyvtártudomány

Fontos szempontnak kell tekintenünk e két tudományág kialakulásának és változásának irányát. Natalia Cecire például kiemeli, hogy a digitális bölcsészet nemcsak a „hagyományos" humán tudományok elméleti alapjaira támaszkodva, tehát kívülről befelé haladva alakult ki, hanem fejlődésének a belülről kifelé történő mozgás is sajátja, hiszen gyakran nem az elmélet, hanem a módszerek állnak középpontjában.[2] A könyvtártudomány viszont ennél egyszerűbb utat járt be azzal, hogy esetében a gyakorlat megelőzte az elméletet, mivel létrejöttét a könyvtárosság gyakorlata alapozta meg. Kialakulását ilyen módon kizárólag a kívülről befelé haladó fejlődés határozta meg.[3]

A könyvtártudomány egyik kiemelkedő teoretikusa, Marcia J. Bates szerint ennek a tudományágnak az a célja, hogy elméleti szinten megismerje és megértse az információ világát és annak emberi tényezőit, továbbá valós szakmai problémák megoldása érdekében az információ szervezése, visszakeresése és terjesztése céljára a gyakorlatban használható eszközöket alakítson ki.[4] Lyn Robinson ehhez hozzáteszi, hogy könyvtártudomány tárgya az információ teljes kommunikációs lánca és annak elemzése. Ez a lánc (és vele a könyvtártudományi oktatás) magába foglalja az információforrások típusait, gyűjtését és gondozását. Része az információ szervezése metaadatok segítségével, az információk és adatok kezelése, az információs viselkedés, valamint az információs műveltség vizsgálata.[5] Nem lényegtelen, hogy a könyvtártudomány számára kiemelkedő szerepe van a taxonómiák, metaadatsémák, programnyelvek, statisztikai technikák és szoftverek alkotta információs infrastruktúrának is.[6] Megítélésem szerint ezek a tartalmak a digitális bölcsészet számára is relevánsak.

2.1. Módszerek és megközelítések

Az alkalmazott módszerek tekintetében egyaránt találunk hasonlóságokat és különbségeket a két terület között. Ahogy arra többek között Axel Bruns felhívja a figyelmünket, a digitális bölcsészet területén dolgozó kutatók matematikai és statisztikai módszereket vesznek kölcsön a számítástudománytól és a természettudományoktól, ami különösen igaz a nagy adatok felhasználásával történő kutatásokra.[7]

[2] Natalia Cecire, „Introduction: Theory and the Virtues of Digital Humanities," *Journal of Digital Humanities* 1, 1. sz. (2011), http://journalofdigitalhumanities.org/1–1/introduction––theory–and–the–virtues–of–digital–humanities–by–natalia–cecire/.

[3] Koltay Tibor, „Library and Information Science and the Digital Humanities: Perceived and Real Strengths and Weaknesses," *Journal of Documentation* 72, 4. sz. (2016): 781–792, 783, https://doi.org/10.1108/jdoc–01–2016–0008.

[4] Marcia J. Bates, „The Invisible Substrate of Information Science," *Journal of the American Society for Information Science* 5, 12. sz. (1999): 1043–1050, 1044, https://doi.org/10.1002/(sici)1097––4571(1999)50:12%3C1043::aid–asi1%3E3.3.co;2–o.

[5] Lyn Robinson, „Information Science: Communication Chain and Domain Analysis," *Journal of Documentation* 65, 4. sz. (2009): 578–591, 582, https://doi.org/10.1108/00220410910970267.

[6] Ma Lia, „Is Information Still Relevant?" *Information Research* 18 (2013), http://InformationR.net/ir/18–3/colis/paperC33.html.

[7] Axel Bruns, „Faster than the Speed of Print: Reconciling 'Big Data' Social Media Analysis and Academic Scholarship," *First Monday* 18, 10. sz. (2013), https://doi.org/10.5210/fm.v18i10.4879.

 DOI 10.31400/dh-hun.2019.2.282

A könyvtártudomány is importál tudást és módszereket más szakterületről, a kvantitatív megközelítéseket azonban többnyire csak kiegészítő jelleggel használja.[8] Több részterülete viszont a humán tudományok módszereit alkalmazza. Példa erre egyik központi kérdésköre, a relevancia, amelynek vizsgálata elképzelhetetlen lenne a nyelvfilozófia és a szemantika eredményeinek felhasználása nélkül.[9]

2.2. A diszciplináris hovatartozás

Ha a kutatás fókuszát tekintjük, akkor érdemes megnéznünk, hogy milyen fordulatok következtek be a könyvtártudomány szemléletében. Ahogy arra Jan Nolin rámutat, ezek a fordulatok többé-kevésbé alapvető változásokban öltöttek testet, viszont nem jelentették a korábbi megközelítésektől való teljes elfordulást. Ilyen volt a történeti fordulat, amely a könyvtártudomány önazonosságának keresését mutatta. A nyelvészeti fordulatot a filozófia diszkurzív, a nyelv funkcióit újraértelmező megközelítése hozta magával.[10] Hozzáteszem, hogy ezek a fordulatok még ma is, nem kis mértékben a humán tudományok episztemológiájához kötik a könyvtártudományt.

Azt, hogy a könyvtártudományt társadalomtudománynak tekintjük,[11] fordulatai közül szociológiai[12] és társadalmi-kognitív paradigmájának[13] előtérbe kerülése erősítette meg. Mindazonáltal a többi tudományághoz fűződő viszonya számos tekintetben tisztázásra vár.[14] Ez azért is van így, mert rendkívül sok és sokféle kérdésre keresi a választ, ami egyaránt tekinthető a gyengeségének és az erősségének.[15] Ennek kapcsán úgy vélem, hogy a könyvtártudomány és a digitális bölcsészet rokonságának egyik fontos jele éppen az ilyen, vélt vagy valós erősségek és gyengeségek megléte.[16]

2.3. A digitális kultúra szerepe

A hasonlóság a kultúrára vonatkozóan is megvan e két tudományterület között. Egyrészt – ahogy azt Michael Buckland megállapítja – nem kétséges, hogy a könyvtár-

[8] Michael Buckland, „What Kind of Science can Information Science Be?" *Journal of the American Society for Information Science and Technology* 63, 1. sz. (2012): 1–7, https://doi.org/10.1002/asi.21656.

[9] James M. Budd, „Relevance: Language, Semantics, Philosophy," *Library Trends* 52, 3. sz. (2004): 447–462.

[10] Jan Nolin, „What's in a Turn?" *Information Research* 12, 4. sz. (2007), http://Informationr.net/ir/12-4/colis/colis11.html.

[11] Michael H. Harris, „The Dialectic of Defeat: Antimonies in Research in Library and Information Science," *Library Trends* 34, 3. sz. (1986): 515–531.

[12] Blaise Cronin, „The Sociological Turn in Information Science," *Journal of Information Science* 34, 4. sz. (2008): 465–475.

[13] Birger Hjørland and Hanne Albrechtsen, „Toward a New Horizon in Information Science: Domain-Analysis," *Journal of the American Society for Information Science*, 46, 6. sz. (1995): 400–425, https://doi.org/10.1002/(sici)1097–4571(199507)46:6%3C400::aid–asi2%3E3.0.co;2–y.

[14] Lyn Robinson and Murat Karamuftuoglu, „The Nature of Information Science: Changing Models," *Information Research* 15, 4. sz. (2010), http://InformationR.net/ir/15–4/colis717.html.

[15] Jan Nolin and Fredrik Åström, „Turning Weakness into Strength: Strategies for Future LIS," *Journal of Documentation* 66, 1. sz. (2010): 7–27, https://doi.org/10.1108/00220411011016344.

[16] Koltay, „Library and Information Science," 782.

tudomány fókuszában a kulturális elkötelezettség áll.[17] Ha pedig elfogadjuk David Berrynek azt a megállapítását, hogy a számítógépes kód a digitális kultúra indexeként szolgálhat,[18] akkor joggal gondolhatjuk, hogy annak is van kulturális aspektusa, hogy a digitális bölcsészet a számítástechnikára támaszkodik.

A kulturális irányultság fontosságát Helle Porsdam szintén megerősíti. Ő úgy látja ugyanis, hogy a digitális bölcsészetnek nemcsak a digitális eszközökkel létrehozott kultúrára kell figyelnie, hanem az is a célja, hogy segítségével megismerjük a számítástechnika alkalmazásának kulturális dimenzióját.[19] Berry szerint ez megköveteli, hogy a kulturális tárgyakat digitális kóddá alakítva tanulmányozzuk, figyelve arra, hogy miként válnak a médium változásai episztemológiai természetűvé, és hogyan alakítják át a szoftverek a tudást információvá. Az (egyre inkább meghatározó) számítógépes kód ugyanis új kommunikatív folyamatokat képes generálni, és a közösségi média egyre nagyobb fontossága magában hordozza az együttműködésre épülő gondolkodás új és izgalmas formáinak a lehetőségét. Ennél fogva a számítástechnikai és adatközpontú témákhoz egyre inkább hozzákapcsolódik a technológia humán értelmezése.[20] Ezt a gondolkodást kell kiegészítenie annak, amit Federica Frabetti úgy fogalmazott meg, hogy elkezdünk a szoftverekre úgy tekinteni, mint az írás és olvasás körébe tartozó problémára.[21]

Ahogy pedig azt John Unsworth kiemeli, a digitális bölcsészet episztemológiai alapkészletének részét képezik a digitális eszközök, amelyeket arra használ, hogy segítségükkel modellezhetővé váljanak a humántudományi adatok, ami jóval több annál, mint amikor a számítógép az írógépet vagy a telefont utánozza.[22] Részben ebből ered, hogy a humán tudományoknak (akár hagyományos formájukban is) meg kell küzdeniük azzal, hogy kialakítsák a technológia használatára vonatkozó elveiket annak érdekében, hogy megmaradjanak a humanista ideálok, viszont megérthessék azt a hatást, amelyet az egyre nagyobb szerepet kapó digitális infrastruktúra gyakorol a tudáslétrehozás rendszerére.[23]

Andrew Dillon szerint a könyvtártudomány képviselőinek célszerű lenne megvizsgálniuk azt, hogy miként tudják pozitív irányba befolyásolni a tág értelemben

[17] Buckland, „What Kind of Science," 4.

[18] David, M. Berry, „The Computational Turn: Thinking about the Digital Humanities," *Culture Machine* 12 (2011): 5, https://culturemachine.net/wp-content/uploads/2019/01/10-Computational-Turn-440-893-1-PB.pdf.

[19] Helle Porsdam, „On Finding the Proper Balance between Qualitative and Quantitative Ways of Doing Research in the Humanities," *DHQ: Digital Humanities Quarterly* 7, 3. sz. (2013): 11, http://www.digitalhumanities.org/dhq/vol/7/3/000167/000167.html.

[20] Berry, „The Computational Turn," 9.

[21] Federica Frabetti, „Rethinking the Digital Humanities in the Context of Originary Technicity," *Culture Machine* 1, 2. sz. (2011): 1–22, https://culturemachine.net/wp-content/uploads/2019/01/1-Rethinking-431-884-1-PB.pdf.

[22] John Unsworth, „What is Humanities Computing, and What is Not?" *Jahrbuch für Computerphilologie* 4, (2002), http://computerphilologie.uni-muenchen.de/jg02/unsworth.html.

[23] Marija Dalbello, „A Genealogy of Digital Humanities," *Journal of Documentation* 67, 3. sz. (2011): 480–506, 482, https://doi.org/10.1108/00220411111124550.

 DOI 10.31400/dh-hun.2019.2.282

vett hálózati infrastruktúra fejlődését.[24] Ez a célkitűzés ráadásul jól illeszkedik a digitális kultúra kritikai tanulmányozásának fentebb említett narratívájához, mivel az *újmédiá*nak az azt használó egyénekre és a tágabb társadalomra gyakorolt hatását vizsgálja.[25]

2.4. Az adatokhoz és az információhoz fűződő viszony

Bár az információ iránti elkötelezettség és érdeklődés eltérő mértékű, továbbá jellege is részben eltér egymástól a digitális bölcsészet és a könyvtártudomány területén,[26] minden más hasonlóságnál erősebb kapcsolatot jelent az, hogy képviselőik számára fontos a szövegben megtestesülő információ léte és annak interpretációja. Ennél fogva különös súlyt kap az – a digitális bölcsészet szempontjából egyébként is alapvető jelentőségű feltételezés –, hogy a szöveget adatokként, az adatokat pedig szövegként értelmezhetjük. Ahogy azt Trevor Owens megfogalmazza, az adatok – mivel van meghatározott célközönségük – értelmezhető és elemezhető szövegeknek tekinthetők.[27] Christof Schöch pedig leszögezi, hogy a humán tudományi adatokat nemcsak digitálisan és szelektíven hozzuk létre, hanem azok az adott vizsgálati tárgy bizonyos tulajdonságainak részlegesen gépi absztrakciói is.[28] Természetesen az ilyen vagy hasonló alapokon nyugvó elképzelések fő pillére az a megközelítés, hogy az adatból meghatározható az információ, az információból a tudás, méghozzá úgy, hogy feltételezzük, hogy ezek a folyamatok az ellenkező irányban is működnek.[29] Joyline Makani érveit követve teljes joggal merülhet fel bennünk, hogy ez a kapcsolatrendszer nem olyan egyszerű, mint ahogy azt az adat, az információ és a tudás összefüggésének hierarchikus modelljei leírják. Az adatok és információk ugyanis kölcsönhatásban állnak egymással, és értéküket az a cél határozza meg, amelynek elérésére felhasználjuk őket.[30] Ezt előlegezi meg Buckland gondolatmenete, amelyben kimondja, hogy az információnak három létformája van.[31] Az első létforma (az információ mint tudás)

[24] Andrew Dillon, „Library and Information Science as a Research Domain: Problems and Prospects," *Information Research* 12, 4. sz. (2007), http://InformationR.net/ir/12–4/colis/colis03.html.

[25] Oya Rieger, „Framing Digital Humanities: The Role of New Media in Humanities Scholarship," *First Monday* 15, 10. sz. (2010), https://doi.org/10.5210/fm.v15i10.3198.

[26] Koltay Tibor, „Könyvtártudomány és digitális bölcsészet: Az információ tudományai?" *Információs Társadalom* 13, 2. sz. (2013): 26–37.

[27] Trevor Owens, „Defining Data for Humanists: Text, Artifact, Digital or Evidence?" *Journal of Digital Humanities* 1, 1. sz. (2011), http://journalofdigitalhumanities.org/1–1/defining–data–for–humanists–by–trevor–owens/.

[28] Christof Schöch, „Big? Smart? Clean? Messy? Data in the Humanities," *Journal of Digital Humanities* 2, 3. sz. (2013): 2–13, http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-Humanities/.

[29] Lin Wang, „Twinning Data Science with Information Science in Schools of Library and Information Science," *Journal of Documentation* 74, 6. sz. (2018): 1243–1257, https://doi.org/10.1108/jd––02–2018–0036.

[30] Joyline Makani, „Knowledge Management, Research Data Management, and University Scholarship: Towards an Integrated Institutional Research Data Management Support-System Framework," *VINE* 45, 3. sz. (2015): 344–359, https://doi.org/10.1108/vine–07–2014–0047.

[31] Michael Buckland, „Information as thing," *Journal of the American Society for Information Science* 42, 5. sz. (1991): 351–360.

azonos az átadott tudással. A második létforma (az információ mint folyamat) tudatállapotunkat módosíthatja. A harmadik létforma az információ mint dolog. Ebben a létformában az információra úgy tekintünk, mint kézzelfogható, rögzített entitásra, amelyet ki tudunk fejezni, le tudunk írni, reprezentációk formájában tükrözni tudunk, vagy fizikailag (jelként) képviselve van. A szöveg ennek a létformának a megtestesülése.

2.5. További közös területek

Golub és Hansson szerint – mivel a könyvtártudomány érdeklődési körébe mindig is beletartoztak az információ szervezésének és terjesztésének problémái, az adatok és az információk közötti rokonság okán e tudományterület egyre gyakrabban foglalkozik az adatok gondozásának és szervezésének kérdéseivel. Az adatok iránti elkötelezettségét erősíti az is, hogy nemcsak az empirikus kutatás eredményeit és a statisztikai elemzések nyersanyagát tekinti adatnak, hanem önálló, saját jogán vizsgálandó kutatási tárgyat is lát benne.[32]

Ugyanakkor mindkét szakterület szakembereinek számolnia kell azzal, hogy a közösségi média fontosságának növekedésével megteremtődik annak lehetősége, hogy az együttműködésre épülő gondolkodás új és izgalmas formái alakuljanak ki. A kérdés viszont az, hogy a szoftverek és kódok valami olyat hoznak-e, ami valódi együttműködést tesz lehetővé, hozzásegítve bennünket olyan, „szuper kritikai" gondolkodás eléréséhez, ami új eszméket, gondolkodási módokat és gyakorlatokat generál.[33] Wendel Piez szerint ezért a digitális bölcsészetnek olyan kritikai attitűdre van szüksége, amely a digitális média tanulmányozásától eljut a média újjáalakításához és újra-feltalálásához,[34] ami meglátásom szerint nincsen másként a könyvtártudomány esetében sem.

Részben a fentiek függvényeként több olyan témát azonosíthatunk, amely a könyvtártudományi kutatás mellett a digitális bölcsészet érdeklődésére is számíthat. Ezek a következők:

- a dokumentumok tartalmi feltárása,[35]
- a digitalizálás és a digitális dokumentumok (szövegek) megőrzése,[36]
- a digitális könyvtárak építése,[37]

[32] Koraljka Golub and Joacim Hansson, „(Big) Data in Library and Information Science: A Brief Overview of Some Important Problem Areas," *Journal of Universal Computer Science* 23, 1. sz. (2017): 1098–1108, 1100.

[33] Berry, „The Computational Turn," 8.

[34] Wendell Piez, „Something Called Digital Humanities," *DHQ: Digital Humanities Quarterly* 2, 1. sz. (2008), http://www.digitalhumanities.org/dhq/vol/2/1/000020/000020.html.

[35] Michael Sperberg-McQueen, „Classification and its Structures," in *A Companion to Digital Humanities*, eds. Raymond George Siemens, Susan Schreibman and John Unsworth (Oxford: Blackwell, 2004), 161–176, https://doi.org/10.1111/b.9781405103213.2004.00017.x.

[36] Marilyn Deegan and Simon Tanner, „Conversion of Primary Sources," in *A Companion to Digital Humanities*, eds. Raymond George Siemens, Susan Schreibman, and John Unsworth (Oxford: Blackwell, 2004), 488–504, https://doi.org/10.1111/b.9781405103213.2004.00035.x.

[37] Howard Besser, „The Past, Present, and Future of Digital Libraries," in *A Companion to Digital Humanities*, eds. Raymond George Siemens, Susan Schreibman and John Unsworth (Oxford: Blackwell, 2004), 557–575, https://doi.org/10.1111/b.9781405103213.2004.00039.x.

 DOI 10.31400/dh-hun.2019.2.282

- a publikációkhoz való nyílt hozzáférés,[38]
- az információ-visszakeresés,[39]
- az eleve digitális dokumentumok,[40]
- a digitalizálás és a digitális megőrzés,[41]
- a nyílt hozzáférés.[42]

Ezt a képet árnyalja a digitális objektumok olyan, a funkcionalitást hangsúlyozó megközelítése, mint az információépítészet, amelynek egyes elemei (így az interfészek és a használhatóság) megjelennek a digitális bölcsészetben,[43] és úgy látom, hogy a könyvtártudomány szempontjából sem érdektelenek.

3. A digitális bölcsészet és a könyvtárak

A digitális bölcsészet és a könyvtárak kapcsolatáról szólva érdemes megjegyeznünk Christine Borgmannak, a tudományos adatokról való interdiszciplináris gondolkodás kiemelkedő képviselőjének és a digitális bölcsészet eszméje támogatójának[44] a kulcsmondatát: „A tudományos kutatáshoz nem több adat kell, hanem a megfelelő adatokra van szükség."[45] Megítélésem szerint ez igaz a digitális bölcsészeti kutatásokra is. Márpedig ha ez így van, akkor a digitális bölcsészet terén dolgozó kutatóknak jó minőségű, gondozott adatokra van szükségük ahhoz, hogy eredményeket érjenek el. Ennek a célnak az elérését nagyban segíthetik a könyvtárak.

Az ilyen irányú együttműködés lehetőségei kapcsán Kiszl Péter és Móring Tibor a könyvtári szférában is eredményesen használható digitális bölcsészeti alkalmazásokat veszi számba.[46] Jelen tanulmányban a kutatók és a könyvtárak közötti együttműködés és a könyvtárak által nyújtott, tág értelemben vett szolgáltatás kereteit, lehetőségeit és példáinak egy részét mutatom be.

[38] Rieger, „Framing digital humanities."

[39] Michael H. Harris, „The Dialectic of Defeat," 523.

[40] Geoffrey Little, „We are all digital humanists now," *The Journal of Academic Librarianship*, 37, 4 sz. (2011): 352–354, https://doi.org/10.1016/j.acalib.2011.04.023.

[41] Deegan and Tanner, „Conversion of Primary Sources," 488.

[42] Matthew G. Kirschenbaum, „So the Colors Cover the Wires: Interface, Aesthetics, and Usability," in *A Companion to Digital Humanities*, eds. Raymond George Siemens, Susan Schreibman, and John Unsworth (Oxford: Blackwell, 2004), 523–542, https://doi.org/10.1002/9780470999875.ch34.

[43] Kirschenbaum, „So the Colors Cover," 523–542.

[44] Christine L. Borgman, „The Digital Future is Now: A Call to Action for the Humanities," *Digital Humanities Quarterly* 3, 4. sz. (2009), http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html/000077.html.

[45] Christine L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World* (Cambridge, MA: MIT Press, 2015), 91–92, https://doi.org/10.1162/leon_r_01170.

[46] Kiszl Péter és Móring Tibor, „Digitális bölcsészet a könyvtár- és információtudományban, 1. rész: A digitális bölcsészet alkalmazásválasztéka," *Tudományos és Műszaki Tájékoztatás* 65, 11. sz. (2018): 574–569, https://doi.org/10.31400/dh–hun.2018.1.228.

3.1. Miért szolgáltatnak a könyvtárak?

Miközben az adatintenzív tudományos kutatás paradigmája a digitális bölcsészet létének előfeltétele, a szolgáltatások egy része véleményem szerint független attól, hogy a digitális bölcsészet vagy más szakterületek kutatóinak kínálják-e őket a könyvtárak, másrészt kétségtelen, hogy a digitális bölcsészeknek is vannak sajátos információs igényeik, amelyek eltérnek más, az adatintenzív tudomány területén megjelenő elvárásoktól. A könyvtárak feladata tehát egyrészt az adatok szervezése és terjesztése, másrészt a kutatási adatok kezeléséhez kötődő szolgáltatások hátterében álló technológiákat is ki kell alakítaniuk.[47] Ezt megkönnyíti, hogy számos feladatot már meglevő ismereteik birtokában tudnak megoldani,[48] és számos hagyományosnak tekinthető készségük hasznosulhat a digitális bölcsészet támogatásában.[49]

Az egyik lehetséges szolgáltatás a kutatók igényeit kielégítő, már létező adatállományok felderítése, mivel az adatok a könyvtárak és az adatokkal foglalkozó munkatársaik gondjaira vannak bízva, és ők azok, akik használatra és elemzések elvégezéséhez rendelkezésre bocsátják azokat.[50] Ezzel cseng egybe, hogy Sennyey Pongrácz a digitális bölcsészet reális működésének kritériumai között említi a létező forrásanyag digitális feldolgozhatóságát.[51]

A könyvtárak küldetése, hogy a kiváló minőségű és megfelelően gondozott adatok iránt megnyilvánuló igényt kielégítsék. Erre akkor lesznek képesek, ha tudják, hogy a kutatók mit gondolnak a könyvtárak nekik nyújtott szolgáltatásairól.[52] Emellett tisztában kell lenniük azzal, hogy saját maguk milyennek látják a szerepüket a kutatóknak nyújtott szolgáltatásokban, továbbá hogyan akarják kutatástámogató feladatukat ellátni.[53] Sok más szakemberrel együtt úgy gondolom, hogy a könyvtárak feladatainak felmérésében és céljaik megvalósításában segítségükre van a munkájukat és szakértelmüket megalapozó könyvtártudomány.

3.2. Példák a szolgáltatásokra és az együttműködésre

Az általános jellemzők után, immár a konkrét feladatokhoz eljutva láthatjuk, hogy viszonylag gyakori szolgáltatás az adatkezelési tervek elkészítéséhez nyújtott segít-

[47] Robin Rice and John Southall, *The Data Librarian's Handbook* (London: Facet Publishing, 2016).

[48] Andrew Cox, „Academic Librarianship as a Data Profession," *Information Today Europe ILI365* 2, (2018): 1–2, https://www.infotoday.eu/Articles/Editorial/Featured-Articles/Academic-librarianship-as-a-data-profession-125376.aspx.

[49] Tim Bryson, Mariam Posner, Alain St, Pierre and Stewart Varner, *Digital Humanities SPEC Kit 326* (Washington, DC: Association of Research Libraries, 2011), https://doi.org/10.29242/spec.326.

[50] Liz Lyon and Eleanor Mattern, „Education for Real-world Data Science Roles (Part 2): A Translational Approach to Curriculum Development," *International Journal of Digital Curation* 1, 6. sz. (2016): 13–26, https://doi.org/10.2218/ijdc.v11i2.417.

[51] Sennyey, *Viták és víziók*, 115.

[52] Barbara Brydges and Kim Clarke, „Is it Time to Re-envision the Role of Academic Librarians in Faculty Research?" *Library Connect* 13 (2015), https://libraryconnect.elsevier.com/articles/2015-07/it-time-re-envision-role-academic-librarians-faculty-research.

[53] Cox, „Academic librarianship," 2.

 DOI 10.31400/dh-hun.2019.2.282

ség.[54] A kutatásfinanszírozó szervezetek egy része előírja ugyanis, hogy az általuk támogatott kutatásokhoz készüljön adatkezelési terv *(Data Management Plan, DMP)*. Ezeknek a terveknek az összeállításában tudnak segíteni a könyvtárosok, amelyhez hasonló készségekre van szükségük, mint amelyekkel akkor élnek, amikor az információk használatával összefüggő (hagyományos) tanácsokkal szolgálnak a felhasználók részére. Az adatkezelési tervek esetében azonban arra is szükség van, hogy ismerjék a finanszírozók előírásait és a helyi adatkezelési folyamatokat. Hozzá kell tennem, hogy viszonylagos elterjedtségük ellenére, jelenleg csak a kutatók kisebb részét érintik az adatkezelési tervekkel kapcsolatos teendők, amely azonban változhat a jövőben. Közben tudjuk, hogy az Amerikai Egyesült Államokban, a *National Endowment for the Humanities* Digitális Bölcsészeti Irodája *(Office of Digital Humanities, ODH)* 2011 óta elvárja, hogy a kutatási pályázatok tartalmazzanak adatkezelési tervet, amelyben a pályázók a kutatás eredményeként keletkező adatok típusára, valamint azok kutatás közbeni és azt követő kezelésére fókuszálnak. Az *ODH* elvárja, hogy a nyertes pályázók terjesszék eredményeiket a tudományos közösség és nagyközönség körében.[55]

Könyvtári feladat lehet az is, hogy a könyvtárosok bibliometriai és alternatív metrikák *(altmetrics)* alapján történő számításokat népszerűsítsenek, sőt ezeket el is végezhetik.[56]

Ahogy arra Robin Rice és John Southall rámutat, a könyvtárosok a kutatókat referensz-interjúk készítésével is segíthetik. Ez a könyvtárak hagyományos tevékenysége, amely azt célozza, hogy munkatársaik megtudják, milyen információra van olvasóiknak szüksége. Ha azonban adatok iránti igényeket akarnak ilyen formában felmérni, az interjú általában több kérdésből áll, mint hagyományos formája, amely publikációk azonosítására irányul. A válaszok várhatóan az adatok kapcsán is hasznosnak bizonyulnak majd, viszont előfordulhat, hogy nem lesznek véglegesnek tekinthetők.[57]

Az adatokra való hivatkozás is olyan terület, ahol a könyvtárosok segíteni tudják a kutatók munkáját. Az adatállományok származásának dokumentálása érdekében pedig metaadatokkal láthatják el az adatállományokat.[58] Végül, de nem utolsósorban, a könyvtárosok a repozitóriumokban elhelyezendő adatállományok kiválasztásában is közreműködhetnek, hiszen vannak ismereteik és tapasztalataik gyűjtemények kialakításában.[59]

A könyvtárosok és a kutatók közös feladata lehet, hogy részt vegyenek a jövő digitális bölcsészeti kutatóinak információs műveltségi oktatásában, ami gyakran a már tapasztalt kutatók és a könyvtárosok együttműködésében valósul meg. Ez az oktatási

[54] Carol Tenopir, Sanna Talja, Wolfram Horstmann, Elina Late, Dane Hughes, Danielle Pollock, Birgit Schmidt, Lynn Baird, Robert Sandusky and Suzie Allard, „Research Data Services in European Academic Research Libraries," *LIBER Quarterly* 27, 1. sz. (2017): 23–44, `https://doi.org/10.18352/lq.10180`.

[55] Alex Poole, „A Greatly Unexplored Area: Digital Curation and Innovation in Digital Humanities," *Journal of the Association for Information Science and Technology* 68, 7. sz. (2017): 1772–1781, `https://doi.org/10.1002/asi.23743`.

[56] Andrew M. Cox and Eddy Verbaan, *Exploring Research Data Management* (London: Facet, 2018).

[57] Rice and Southall, *The Data Librarian's Handbook*, 75.

[58] Cox, „*Academic librarianship,*" 1.

[59] Cox and Verbaan, *Exploring Research Data,* 150.

tevékenység nem teljesen új és – bár sokan vitatják létjogosultságát – folyamatosan fejlődik.[60] Az információs műveltség mellett (sok tekintetben annak részeként) ott van az adatműveltség is, amelyet készségek és tudásbázis együtteseként határozhatunk meg. Célja, hogy lehetővé tegye számunkra az adatok információvá és a gyakorlatban használható tudássá alakítását olyan módon, amely képessé tesz az adatok elérésére, értelmezésére, kritikai értékelésére, kezelésére és etikus használatára.[61]

Számolnunk kell azonban azzal, hogy sok könyvtáros egyelőre nincsen kellő mértékben felkészülve arra, hogy kielégítse a digitális bölcsészeti kutatások által támasztott igényeket. Ahogy arra többen is rámutatnak, a könyvtárosok gyakran a napról napra felmerülő, egyedi igények kielégítésére fókuszálnak, továbbá előnyben részesítik a kezdő kutatók támogatását, és főként a kutatás kezdeti fázisában működnek közre, [62] ami akadálya lehet annak, hogy a kutatók tágabb körére is kiterjesszék szolgáltatásaikat.[63]

Bár a digitális bölcsészet szakirodalmában gyakran találkozunk azzal a gondolattal, hogy az oktatásában és a kutatásban fontos szerepe és jótékony hatása van az együttműködésnek, megvalósítása gyakran ütközik akadályokba. Mindazonáltal Bethany Nowviskie joggal hangsúlyozza, hogy számos olyan könyvtáros van, aki be tud illeszkedni a digitális bölcsészeti kutatás és oktatás folyamatába.[64] Közülük sokan arra is késztetést éreznek, hogy teljes jogú partnerként vegyenek részt digitális bölcsészeti projektekben. Itt ismét érdemes visszatérnünk a kutatási adatok kezelésében történő olyan jellegű részvételre, amely egyelőre még ritka a humán tudományi (így a digitális bölcsészeti) kutatások esetében is. Ennek egyik terepe lehet az, hogy a könyvtárosok információforrásokat biztosítanak a kutatáshoz, és egyúttal oktatják is a hallgatókat ezeknek a forrásoknak a használatára.[65] John Russel és Merinda Kaye Hensley ugyanakkor felhívja a figyelmet arra, hogy a könyvtárosok az oktatási feladatköröket csak nemrégiben kezdték el feltérképezni. Rámutatnak arra is, hogy az oktatás a digitális eszközök bemutatására korlátozódik, miközben a könyvtárosoknak egyre inkább arra volna igénye, hogy ne csak a szoftverek sajátosságait ismertessék meg a hallgatókkal, hanem a digitális bölcsészet tágabb kontextusát is megmutassák, és

[60] Ying Zhang, Shu Liu and Emilee Mathews, „Convergence of Digital Humanities and Digital Libraries," *Library Management* 36, 4–5. sz. (2015): 362–377, https://doi.org/10.1108/lm—09—2014——0116.

[61] Koltay Tibor, „Data Literacy: in Search of a Name and Identity," *Journal of Documentation* 71, 2. sz. (2015): 401–415, https://doi.org/10.1108/jd—02—2014—0026.

[62] Ixchel M. Faniel and Lynn Silipigni Connaway, „Librarians' Perspectives on the Factors Influencing Research Data Management Programs," *College and Research Libraries* 79, 1. sz. (2018): 100–119, https://doi.org/10.1108/jd—02—2014—0026.

[63] Matt Burton and Liz Lyon, „Data Science in Libraries," *Bulletin of the Association for Information Science and Technology* 43, 4. sz. (2017): 33–35.

[64] Bethany Nowviskie, „Skunks in the Library: A Path to Production for Scholarly R&D," *Journal of Library Administration* 53, 1. sz. (2013): 53–66, https://doi.org/10.1080/01930826.2013.756698.

[65] Janet Hauck, „From Service to Synergy: Embedding Librarians in a Digital Humanities Project," *College and Undergraduate Libraries* 24, 2–4. sz. (2017): 434–451, https://doi.org/10.1080/10691316.2017.1341357.

 DOI 10.31400/dh-hun.2019.2.282

tanulási élményeket is nyújtsanak nekik, valamint a digitális módszerek és a források kritikai megközelítését állítsák a középpontba.[66]

Thoughts on the Relationship between Digital Humanities, Library and Information Science and Libraries

Libraries and librarians can support digital humanists not only by having expertise in identifying objectives and working out ways of realization, informed by library and information science (LIS) but because both the digital humanities and LIS have several themes and issues in common. LIS aims to explore and understand the world of information and its human factors, and the latter is in the focus of digital humanists' research as well. Besides showcasing these common features, examples of services offered to digital humanists (mainly) in Anglo-American academic libraries are presented in the paper.

Keywords:
library and information science, digital humanities research, academic libraries, library services, cooperation

[66] John E. Russel and Merinda Kaye Hensley, „Beyond Buttonology: Digital Humanities, Digital Pedagogy, and the ACRL Framework," *College and Research Libraries News* 78, 11. sz. (2017): 588–600, https://doi.org/10.5860/crln.78.11.588.

Kiss Margit

Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet

kiss.margit@btk.mta.hu

Stilometriai elemzés lehetőségei magyar történeti szövegkorpuszon

Tanulmányomban magyar nyelvű történeti szövegek számítógépes elemzésének egy olyan lehetőségével foglalkozom, amely ötvözi a nyelv- és irodalomtudomány, az informatika és a statisztika eredményeit. A szerzőségi, illetve a stilometriai vizsgálat bár nem új keletű az irodalmi szövegelemzések esetében, módszertanát tekintve folyamatosan formálódik, megújul. Munkámban áttekintem e szövegelemzési módszer jellemzőit és alkalmazási lehetőségeit, majd esettanulmányként különböző elemzéseket mutatok be Mikes Kelemen művei alapján. Stilometriai módszerekkel vizsgálom az életműben a saját szerzőségű művek és a fordítások kapcsolatát, valamint a művek tematikai elkülönülését. Végezetül bemutatom, hogy a digitális írói szótár alkalmazása – mint történeti szöveget normalizáló eszköz – hogyan javíthatja ezeknek az elemzőmódszereknek a hatékonyságát.

Kulcsszavak:
szerzőségi vizsgálat, stilometria, digitális írói szótár, Mikes Kelemen

> A nyelv csodálatossága abban rejlik, hogy bár közös forrásból táplálkozik, mi mind valami egyedit hozunk létre belőle. A számítógépes elemzés lehetővé teszi, hogy sokkal pontosabban nyomozzunk a lexikai elemek után, mintha egyszerűen csak a puszta intuíciónkra hagyatkoznánk.[1]

1. Bevezetés

A nagyméretű szövegkorpuszokat számítógép támogatásával elemző kutatók manapság egyre több módszer közül választhatnak, s olyan kérdésekre adhatnak választ, amelyekre korábban manuális módszerek felhasználásával még nem vagy csak jelentős időráfordítással volt lehetőség. Az egyre nagyobb méreteket öltő digitalizálás

[1] Hugh Craig, a University of Newcastle professzor emeritusának egyetemi weboldalán megjelent összegzés. Hugh Craig, „Figures of Speech," hozzáférés: 2019.02.20, `https://www.newcastle.edu.au/profile/hugh-craig`. (Ford. tőlem.)

mellett arra is érdemes hangsúlyt fektetni, hogy akik korpusz alapú szövegvizsgálatokat végeznek, megismerjék, alkalmazzák, majd továbbfejlesszék az elemzőeljárásokat.[2] A tanulmány célja kettős: a modern nemzetközi kutatások tükrében a nyelv-, az irodalomtudomány, az informatika, valamint a statisztika eredményeit ötvöző, a magyar nyelvű szövegek vizsgálatában kevéssé elterjedt szerzőségi, illetve stilometriai elemzőmódszert mutatom be. Másfelől arra a kérdésre keresem a választ, hogy vajon a magyar történeti szövegek elemzésében hogyan alkalmazhatjuk ezeket a statisztikai alapú módszereket, és hogy miként növelhetjük az elemzések hatékonyságát. Ehhez konkrét esettanulmányokat mutatok be Mikes Kelemen életművének különböző szempontok alapján történő stilometriai elemzésével.

2. Előzmények

2.1. Szerzőségi vizsgálatok

A vitatott vagy bizonytalan szerzőség megállapításával kapcsolatos vizsgálatok éppoly régre nyúlnak vissza, mint amióta az írás létezik. A szerzőségi vizsgálatokat és alakulásukat kutató Hugh Craig megjegyzi, hogy a *Biblia,* a homéroszi művek vagy Shakespeare munkái még egy olyan időszakban születtek, amikor a szerzői homogeneitás nem volt különösebben fontos szempont.[3] Későbbi generációk mégis jelentőséget tulajdonítottak annak, hogy szerzőségi szempontból is megvizsgálják ezeket a szövegeket. A bekövetkezett szemléletváltás a reneszánszra tehető a szövegek komparatív vizsgálatának lehetőségével, a nyelvi és textológiai diszciplínák alkalmazásával. Egyik leghíresebb példa erre Lorenzo Valla 15. századbeli humanista munkája, aki filológiai módszerekkel bizonyította be, hogy a *Donatio Constantini* adománylevél hamisítvány.[4] Azóta számos kétes vagy bizonytalan szerzőségű művet tulajdonítottak valamely szerzőnek, vagy zártak ki egy adott szerzőség alól, de nem kevés szöveg maradt anonim vagy legalábbis vitatott szerzőségű.

A szerzőségi vizsgálatok hagyományos megközelítésben a filológia, nyelv- és irodalomtörténet, paleográfia, kodikológia, történettudomány és az igazságügy egyes területeit érintik, de idővel a diszciplína nem hagyományos eljárásokkal is kiegészült, úgy mint a statisztikai elemzőmódszerek.[5] A szerzőségi vizsgálatok esetében a statisztika alkalmazása ugyan nem nevezhető tradicionálisnak, ugyanakkor a statisztikára támaszkodó szövegelemzés nem új fejlemény.[6] 1851-ben Augustus de Morgan egy

[2] Mészáros Tamás, „Mit nyújthat a modern informatika az irodalomtudomány számára?" *Magyar Tudomány,* 11. sz. (2016): 1310–1315, http://www.matud.iif.hu/2016/11/06.htm.

[3] Hugh Craig, „Stylistic Analysis and Authorship Studies" in *A Companion to Digital Humanities,* eds. Susan Schreibman, Ray Siemens and John Unsworth (Oxford: Blackwell Publishing, 2007), 282, https://doi.org/10.1002/9780470999875.

[4] Harold Love, *Attributing Authorship* (Cambridge: Cambridge University Press, 2002), 18–19, https://doi.org/10.1017/cbo9780511483165.003; Christopher B.Coleman, ed, trans., *The Treatise of Lorenzo Valla on the Donation of Constantine. Text and Translation into English* (New Haven: Yale University Press, 1922), 131–133.

[5] Maciej Eder, „Style-Markers in Authorship Attribution: A Cross-Language Study of the Authorial Fingerprint," *Studies of Polish Linguistics* 1 (2011): 100.

[6] A jelen tanulmány szempontjából releváns mérföldkövekre vonatkozóan felhasznált és áttekintő összegzést adó irodalom: David Holmes and Judit Kardos, „Who Was the Author? An Introduction to

 DOI 10.31400/dh-hun.2019.2.336

barátjának írott levelében arról a megfigyeléséről számol be, hogy a szavak hosszúságának meghatározó szerepe van a szerzőség megállapításával kapcsolatban.[7] Thomas Mendenhall amerikai fizikus az 1880-as években az írói stílus kvantitatív elemzésével foglalkozott, elsősorban angol szerzők munkái alapján.[8] Évtizedekkel később George Udny Yule és George Zipf meghatározó eredményeket ért el az elemzésben alkalmazható szövegjegyek felkutatásában.[9] Az 1960-as években Frederick Mosteller és David Wallace analízise már megnyitotta az utat a modern, digitális kor stilometriája felé: munkáik úttörő jelentőségűvé váltak az irodalmi szövegek szerzőségi vizsgálatainak tekintetében.[10] A *Federalist Papers* 1787 és 1788 között 85 politikai esszét publikált, amelyben a szavazókat az Egyesült Államok számára készített alkotmány jóváhagyásáról igyekeztek meggyőzni. Az esszéket mind „Publius" névvel jegyezték, de azt azért lehetett tudni, hogy Alexander Hamilton, James Madison, illetve John Jay írhatta őket. Több nyelvi megkülönböztető jegyet, valamint valószínűségi modelleket is felhasználtak a különösen nehéznek mutatkozó szerzőségi probléma miatt, amelyet a stílus- és a politikai tartalombeli hasonlóság is nehezített. Frederick Mosteller és David Wallace mind a tizenkét vitatott írást Madisonnak tulajdonította, így a kapott eredmény lényegében összhangban állt a történészkutatók eredményeivel. Az 1980–1990-as években John Burrows jelentős eredményeket ért el új elemzési eljárások kialakításával, amelynek során a megkülönböztető jegyek közül a funkciószavak elemzésére támaszkodott. Burrows több szerzőt és eltérő műfajú munkákat elemzett, például Austent, a Brontë testvéreket, Scottot és Byront.[11] Lexikális szintű elemzések végzése során a Burrows-módszer ma is elterjedt.[12] A számítógép térhódítása a bölcsészettudományokban új szakaszt nyitott a szerzőségi vizsgálatok területén is a szövegek stilisztikai jegyeinek a mérésével, valamint az eredmények összevethetősége és értékelése terén, ugyanakkor

Stylometry," *Chance* 16, 2. sz. (2003): 5–8, http://doi.org/10.1080/09332480.2003.10554842; David Holmes, „The Evolution of Stylometry in Humanities Scholarship," *Literary and Linguistic Computing* 13, 3. sz. (1998): 111–117, https://doi.org/10.1093/llc/13.3.111; Harold Love, *Attributing Authorship* (Cambridge: Cambridge University Press, 2002), https://doi.org/10.1017/cbo9780511483165.003.

7 R. D. Lord, „Studies in the History of Probability and Statistics. VIII. de Morgan and the Statistical Study of Literary Style," *Biometrika* 45, 1–2. sz. (1958): 282, https://doi.org/10.1093/biomet/45.1–2.282.

8 Thomas Corwin Mendenhall, „The Characteristic Curves of Composition," *Science* 9, 214. sz. (1887): 237–249, https://doi.org/10.1126/science.ns–9.214s.237.

9 Udny Yule, *The Statistical Study of Literary Vocabulary* (Cambridge, Cambridge University Press, 1944); George Kingsley Zipf, *Selected Studies of the Principle of Relative Frequency in Language* (Cambridge: Harvard University Press, 1932), https://doi.org/10.4159/harvard.9780674434929.

10 Frederick Mosteller and David Wallace, *Inference and Disputed Authorship: The Federalist.* Reprinted With a New Introduction by John Nerbonne (Stanford: CSLI Publications, 2007 [1964]).

11 John Burrows, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (Oxford: Clarendon Press, 1987).

12 John Burrows, „'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing* 17, 3. sz. (2002): 267–287, https://doi.org/10.1093/llc/17.3.267; David Hoover, „Testing Burrows's Delta," *Literary and Linguistic Computing* 19, 4. sz. (2004): 453–475, https://doi.org/10.1093/llc/19.4.453.

azt is el kell ismerni, hogy a stilometria nemritkán heves szakmai viták kereszttüzébe kerül.[13]

2.2. Stilometria

A stilometria szó és a diszciplína megalkotója Wincenty Lutosławsky.[14] Ő az új módszert Platón dialógusainak a kronologizálásához alkalmazta, amellyel a filozófus eszmerendszerének az értelmezéséhez nyújtott újfajta segédletet. Ma a stilometria megnevezés a stílus statisztikai alapú vizsgálatát jelenti, a szerzőség statisztikai szempontú, nyelvészeti és statisztikai feltevéseken alapuló megközelítését *nem hagyományos szerzőségi vizsgálat*nak (*non-traditional authorship attribution,* ford. tőlem) nevezik.[15] Mindkét esetben az a kérdés áll a középpontban, hogy melyek azok a nyelvi tényezők, amelyek meghatározók a szerzői művekkel kapcsolatban. Az irodalmi nyelv és stílus statisztikai alapú elemzésének nem az a célja, hogy felforgassa a hagyományos elemzések során alkalmazott eszközkészletet, hanem hogy kiegészítse, komplexebbé tegye a hagyományos módszerekkel végzett vizsgálatokat a kétséges jelenségeket illetően. Minden szerzőnek van egy olyan sajátos stílusa, ami állandó, és olyan jegyeket tartalmaz, amely mennyiségileg is meghatározható, ezáltal megkülönböztető funkcióval rendelkezik, így bizonyos nyelvi jellemzők (szókészlettani gazdagság, kollokációk, sajátos szintaktikai jellemzők, szókörnyezet) statisztikai eszközökkel mérhetők. Az a cél, hogy fel lehessen tárni ezeket a szerzői megkülönböztető jegyeket, különösen azokat, amelyek a szoros olvasás során észrevétlenek maradnak. Kísérletek azt bizonyítják, hogy ezek az emberi olvasás során rejtve maradó rögzült minták a stílusparódiák vagy az álnéven írt munkák szerzőit is leleplezhetik azzal, hogy a saját nyelvezetük ujjlenyomatait hordozzák magukon.[16] A szövegelemzés során a számítógépes stilisztika tendenciákkal dolgozik. A tendenciák jobban megfigyelhetők az olyan összetett jelenségek mögött, amelyekhez az emberi feldolgozó- és felfogóképesség már nem elegendő. Azokon a területeken nyújt segítséget, amelyeken több szempontú, átfogó összehasonlítások szükségeltetnek: a nyelvi modellek vizsgálatakor a szövegalkotás, kifejezésmód egyéni jellemzőinek a feltárásában úgy, hogy az egyént meghatározó jegyek kiszűrésére törekszik.

David Holmes és Judit Kardos rámutat arra, hogy a modern stilometria a kezdetek óta sokat változott a számítógép nyújtotta lehetőségek és a mesterséges intelligencia hatására, amely a meghatározó stílusjegyek felismerésében is szerepet játszhat.[17] Rámutatnak továbbá, hogy a gépi tanulás sikeresen alkalmazható e területen. A *neurális*

[13] M. W. A. Smith, „Shakespeare, Stylometry and »Sir Thomas More«," *Studies in Philology* 89, 4. sz. (1992): 434–444.

[14] Lutosławski Wincenty, *The Origin and Growth of Plato's Logic: With an Account of Plato's Style and of the Chronology of his Writings* (London: Longmans, 1897), `https://archive.org/details/originandgrowth00lutogoog/page/n44`.

[15] Eder, „Style-Markers in Authorship," 100–101.

[16] Hugh Craig, „Stylistic Analysis," 285; John Burrows, „I Lisp'd in Numbers: Fielding, Richardson and the Appraisal of Statistical Evidence," *The Scriblerian,* 33 (1991): 234–241. J. K. Rowling Cuckoo's Calling című regényének szerzőazonosítása Patrick Juola, „The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions," *Digital Scholarship in the Humanities* 30, 1. sz. (2015): 100–113, `https://doi.org/10.1093/llc/fqv040`.

[17] Holmes and Kardos, „Who Was the Author," 5–8.

 DOI 10.31400/dh-hun.2019.2.336

hálók segítségével tökéletesíthetjük az elemzést, amely azáltal javítja a módszer működését, hogy maga próbál olyan tulajdonságot felfedezni, amely az általunk megadottat tökéletesíti. A tanítófolyamat számos előnye mellett a hátránya az, hogy nagy mennyiségű adat szükséges hozzá. A *genetikus algoritmus* a tanítókorpuszon kalibrálódik evolúciós jelleggel, és a stilometriai vizsgálatokban a meglévő szabályok közül a legadekvátabb megkülönböztető funkció megtalálását segíti. Úgy vélik, vitatott szerzőség esetében nagyon jó eredménnyel alkalmazható, ha elegendő adat áll rendelkezésre a mintatanuláshoz. 1993–1994-ben Robert Matthews és Tom Merriam alkalmazta a módszert sikerrel: Shakespeare és Fletcher műveiből tanítókorpuszt hoztak létre, majd a *The Two Noble Kinsmen* című műben vizsgálták a két szerző kollaborációját.[18]

2.2.1. Mire alkalmazható a stilometria?

Az, hogy szövegeket a lexikális jegyeik alapján mérünk és hasonlítunk össze, lehetővé teszi a vizsgált szövegek közti azonosságok és különbségek értékelését. A vizsgálati szempontok vonatkozhatnak anonim vagy vitatott szerzőségű szövegek azonosításának a támogatására,[19] de akár egy szerzői munkásságon belül a nyelvezet, a szövegformálás változásának a feltárására is, amely segítséget nyújthat az életmű korszakolásában.[20] Elemezhetünk csoporthoz való tartozást: férfi és női szerzők munkái közti különbséget,[21] műfaji jelleget,[22] nyelvi szempontból is megmutatkozó hatást, előzményt, inspirációt[23] stb.

Ezek az elemzési metódusok jellemzően nem önmagukban állnak, sőt magukban alkalmazva félre is vezethetik az elemzőt. Például e vizsgálatok eredményeképpen ma úgy vélik, hogy Shakespeare-nek nem volt átlagon felüli gazdagságú szókészlete, az ő kivételessége sokkal inkább abban rejlik, hogy egyedülálló módon használta az átlagos, hétköznapi szavakat. A stilometriai módszerek bevonásának köszönhetően ma már valószínűsíthető, hogy a *VI. Henrik* című dráma egyes részeiben Marlowe is közreműködött.

[18] Uo., 5–8. Robert Matthews and Tom Merriam, "Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher," *Literary and Linguistic Computing* 8, 4. sz. (1993): 203–209, https://doi.org/10.1093/llc/8.4.203; Tom Merriam and Robert Matthews, "Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe," *Literary and Linguistic Computing* 9, 1. sz. (1994): 1–6, https://doi.org/10.1093/llc/9.1.1.

[19] Ward E. Y. Elliott and Robert J. Valenza, „Two Tough Nuts to Crack: Did Shakespeare Write the 'Shakespeare' Portions of Sir Thomas More and Edward III? Part I," *Literary and Linguistic Computing* 25, 1. sz. (2010): 67–83, https://doi.org/10.1093/llc/fqp029; Ward E. Y. Elliott and Robert J. Valenza, „Two Tough Nuts to Crack: Did Shakespeare Write the 'Shakespeare' Portions of Sir Thomas More and Edward III? Part II: Conclusion," *Literary and Linguistic Computing* 25, 2. sz. (2010): 165–177, https://doi.org/10.1093/llc/fqp029.

[20] Dirk Van Hulle and Mike Kestemont, „Periodizing Samuel Beckett's Works: A Stylochronometric Approach," Style 6, 2. sz. (2016), 172–202, https://doi.org/10.5325/style.50.2.0172.

[21] Sean G. Weidman and James O'Sullivan, „The Limits of Distinctive Words: Re-evaluating Literature's Gender Marker Debate," *Digital Scholarship in the Humanities* 33, 2. sz. (2018): 374–390, https://doi.org/10.1093/llc/fqx017.

[22] Alexandre Sotov, „Lexical Diversity in a Literary Genre: A Corpus Study of the *Ṛgveda*," *Literary and Linguistic Computing* 24, 4. sz. (2009), 435–447, https://doi.org/10.1093/llc/fqn044.

[23] Regula Hohl Trillini and Sixta Quassdorf, „A 'Key to all Quotations'? A Corpus-Based Parameter Model of Intertextuality," *Literary and Linguistic Computing* 25, 3. sz. (2010), 269–286, https://doi.org/10.1093/llc/fqq003.

ködött.[24] Bár alapvetően az angol nyelvű munkákra és a klasszikus művek vizsgálatára koncentrálódnak a számítógépes szerzőségi elemzővizsgálatok,[25] az utóbbi időszakban más nyelvekre és szövegtípusokra is alkalmazzák őket.[26]

3. Szerzői életmű stilometriai vizsgálata

Ha a szövegek közti eltéréseket kutatjuk, akkor nemcsak az egyes szerzők közti különbözőségeket vizsgálhatjuk, hanem a szerzői életművön belüli váltásokat is nyomon követhetjük. Ez esetben fontos látnunk, hogy a szerzők egymás közti kifejezésmódbeli különbözősége és a szerzői életmű alakulása – bár e két típus közel sem egyforma mértékben – szövegstatisztikai szempontból eltéréseket rejt. Mérhető különbségek nemcsak szerzők között lehetnek, hanem egy életpálya különböző szakaszai között is vizsgálható a nyelvezet változása, amelynek módszeres vizsgálata különféle értelmezői keretek kialakításában is segítséget nyújthat, így például a szerzői életművek szakaszolásában.[27] Az írói-költői nyelvezet alakulásának a vizsgálata során a *Does "Late Style" Exist? New Stylometric Approaches to Variation in Single-Author Corpora* című tanulmány[28] szerzője arra az eredményre jutott, hogy nem az egyes szerzők kései korszakának a beszédmódjai térnek el jelentősen a megelőzőekétől, hanem éppen a korai írói életpálya különül el markánsan a későbbi alkotói szakaszoktól. Az önálló szerzői korpusz vizsgálata az alkotásmód alakulása tekintetében egy lehetséges út a szépirodalmi szövegek stilometriai elemzésében. Különösen azokban az esetekben hatékony eszköz, amelyekben jelentős mennyiségű szöveg áll rendelkezésre az életmű adekvátabb megértése érdekében.

Magyar nyelvű szépirodalmi szövegek vizsgálatában nem általánosan elterjedt gyakorlat a stilometriai módszertan. A magyar történeti szövegek számítógépes elemzése különösen nehézségekkel terhelt feladat a nyelv standardizálatlansága és a gépi elemzés szabályelvűsége miatt. Arra voltunk kíváncsiak, hogy a stilometriai módszerek vajon ezzel együtt is támogatást nyújtanak-e a szövegvizsgálatokban, s a magyar nyelvű történeti szövegek vizsgálatához alkalmazható-e megbízható eredménnyel ez a

[24] Hugh Craig, „Ignore the Doubters: Here's Why Christopher Marlowe Co-wrote Shakespeare's Henry VI," *The Conversation,* 2016. nov. 9, `https://theconversation.com/ignore-the-doubters-heres-why-christopher-marlowe-co-wrote-shakespeares-henry-vi-68229`; Hugh Craig and Arthur F. Kinney, eds., *Shakespeare, Computers and the Mystery of Authorship* (Cambridge: Cambridge University Press, 2009), `https://doi.org/10.1017/cbo9780511605437`.

[25] Vö. 6. jegyzet, különösen Harold Love, *Attributing Authorship* (Cambridge: Cambridge University Press, 2002), `https://doi.org/10.1017/cbo9780511483165.003`.

[26] Érdekes kísérlet a lengyel nyelv történeti korszakolásának vizsgálatára: Maciej Eder and Rafal L. Górski, „Historical Linguistics' New Toys, or Stylometry Applied to the Study of Language Change" in *Digital Humanities 2016: Conference Abstracts,* eds. Maciej Eder and Jan Rybicki (Krakow: Jagellonian University & Pedagogical University, 2016), 182–184.

[27] Van Hulle and Kestemont, „Periodizing Samuel Beckett's Works," 172–202.

[28] Jonathan Pearce Reeve, „Does »Late Style« Exist? New Stylometric Approaches to Variation in Single-Author Corpora" in *Digital Humanities 2018, DH 2018, Book of Abstracts,* eds. Jonathan Girón Palau and Isabel Galina Russell (Mexico City: El Colegio de México, UNAM, and RedHD, June 26–29, 2018), 478–480.

 DOI 10.31400/dh-hun.2019.2.336

módszer. Kísérletképpen ismert szerzőségű szövegeken, a Mikes-korpuszon[29] végeztünk elemzéseket.[30] Előtte azonban érdemes áttekinteni, hogy milyen elemekből áll egy stilometriai elemzés.

3.1. Mit mérjünk?

Lutoslawsky 1897-ben azt írta, hogyha a kézírás meghatározza az írója személyét, akkor az egyéni stílus még ennél is személyesebb és jellemzőbb.[31] Maciej Eder szerint a mai stilometriai módszereket alkalmazók messze állnak ettől a határozottságtól, mégis úgy vélik, hogy az írás folyamatára hatással van a tudattalan.[32] A legfontosabb feladat kinyomozni az erről árulkodó jegyet, a „szerzői ujjlenyomatot" a különféle nyelvi (lexikális, morfológiai, szintaktikai) jellemzők közül. Arra a kérdésre kell választ adni, hogy melyik az a nyelvi jelenség, amely mérhető az egyes szerzői szövegekben a szerzői ujjlenyomat meghatározása érdekében. Úgy véli, hogy a sikeres vizsgálathoz minél több egyedi stíluselem, ún. stílusmarker[33] meghatározása a cél. Hogy a stilisztikai egyéni jellemzők kialakítása során azonban mi a közös és mi az egyedi a nyelvben, nem teljesen magától értetődő. Véleménye szerint a legjobb stílusmarkerek azok, amelyek szabad szemmel felfedezhetetlenek, így a szerzői kontrollon túlmutatnak, és az utánzás sem fog rajtuk. Bár idővel egy szerző stílusa, kifejezésmódja változhat, nem különböznek olyan meghatározó mértékben egymástól a saját szövegeik, mintha más szerzőkhöz hasonlítanánk őket. Hugh Craig például rámutatott, hogy a korai Henry James és a kései Henry James is különböző, de nem annyira eltérő, mint Henry James és Thomas Hardy.[34] Burrows pedig egy vizsgálatában rávilágított arra, hogy Henry Fielding Samuel Richardson ellenében álnéven írt stílusparódiája sokkal közelebb maradt a saját stílusára jellemző nyelvi elemekhez, mint a kifigurázandó szerzőéhez.[35]

A stílusmarkerek változatossága legalább olyan gazdag, mint a története. David Holmes *Authorship Attribution* című írása[36] kiválóan összefoglalja a különféle stílusmarkerek alkalmazásával elért eredményeket, és egyben feltárja gyenge pontjaikat is. A tudomány jelen állása szerint a stilometriában a nyelvi változásnak nem a kevésbé, hanem éppen a jobban ellenálló szókészleti elemek vizsgálata ígérkezik eredményesebbnek, mert ezek mutatnak rá az egyéni kifejezésmód rögzült formáira. A lexikai szint mára meghatározóvá vált, ennek mérhetővé tételére számos különféle statisztikai

[29] Mikes Kelemen, *Összes művei,* s. a. r. Hopp Lajos (Budapest: Akadémiai Kiadó, 1966–1988). Elektronikus verzió: Magyar Elektronikus Könyvtár, 2011, `http://mek.oszk.hu/09000/09000/`. A kísérlethez felhasznált szövegkorpusz a cikk online mellékletében megtalálható (vö. 53. jegyzet): `https://doi.org/10.31400/dh–hun.2019.2.336`.

[30] A stilometriai kísérletek futtatásában Dobi Jan Sándor hallgató (BME VIK) és Mészáros Tamás egyetemi oktató (BME VIK) volt segítségemre.

[31] Eder, „Style-Markers in Authorship," 103 alapján Lutosławski, "The Origin and Growth of Plato's," 66.

[32] Eder, „Style-Markers in Authorship," 103.

[33] A kifejezés az angol terminológia alapján történő tükörfordítás tőlem. A szót 'stílust jelölő' értelemben használom.

[34] Craig, „Stylistic Analysis," 285.

[35] Burrows, „I Lisp'd in Numbers," 234–241.

[36] David Holmes, „Authorship Attribution," *Computers and the Humanities* 28, 2. sz. (1994): 87–106, `https://doi.org/10.1007/bf01830689`.

módszer született.[37] Eder kutatása[38] rámutat, hogy a korszerű vizsgálatokban a legszélesebb körben elterjedt a minimum 100 leggyakoribb szó elemzése (MFW, min. 100), ezt követi a mondathossz, a szóhosszúság, a hangsúlyos és hangsúlytalan szótagok váltakozása, a szókészlet gazdagsága, a leggyakoribb funkciószavak, a központozás, a kollokációk, bizonyos betűsorozatok gyakorisága és a szóbigramok vizsgálata.[39]

Jelen kutatások arra is felhívják a figyelmet, hogy a stílusmarkerek nem tekinthetők teljesen nyelvfüggetlennek.[40] Különbözőségük a nyelvtípusok közti különbségből is adódik. Egyre több vizsgálat irányul e nyelvspecifikus jegyek feltárására.[41] Hogy végül ténylegesen egy adott szöveg elemzéséhez melyik marker válik megkülönböztetővé, az erősen függ magától a korpusztól. Grieve a *Quantitative Authorship Attribution: An Evaluation of Techniques* című tanulmányában[42] harminckilenc szerzőségi módszert hasonlít össze, hogy választ kapjon arra a kérdésre, melyik lehet a leghasznosabb a szerzőség megállapításához. Ismert szerzőségű szövegeken hasonlította össze a különféle lehetőségeket, és arra az eredményre jutott, hogy az algoritmusok kombinációja meggyőző eredményt nyújt a megfelelő stílusmarker megtaláláshoz, de még a valószínűség megfogalmazásához is több módszer együttes alkalmazását tartja indokoltnak.

3.2. Hogyan mérjünk?

A nyelv vizsgálatára alkalmazott statisztikai technikáknak egyik csoportja a szövegkorpuszból körültekintően kiválasztott egyetlen jelenségre fókuszál, mint például a szókészlet gazdagsága, különféle indexek stb. A másik csoport nagy mennyiségű jellemzőt vizsgál, ezek a multidimenzionális statisztikai módszerek, amelyek finomabb különbségek feltárására is alkalmasak.[43] Ezek lényege, hogy a tulajdonságok sokdimenziós terében helyezik el a vizsgált szövegeket. Ilyen például a *klaszteranalízis,* amely egy olyan csoportosító eljárás, amellyel elemeket homogén csoportokba ren-

[37] Holmes, „Authorship Attribution," 87–98.

[38] Eder, „Style-Markers in Authorship," 103.

[39] David Holmes, „The Evolution of Stylometry," 111–117; David L. Hoover, „Frequent Word Sequences and Statistical Stylistics," *Literary and Linguistic Computing* 17, 2. sz. (2002): 157–180, https://doi.org/10.1093/llc/17.2.157; Juan-Pablo Posadas-Duran, Grigori Sidorov and Ildar Batyrshin, „Complete Syntactic N-grams as Style Markers for Authorship Attribution," in *Human-Inspired Computing and Its Applications,* MICAI 2014, Lecture Notes in Computer Science, vol. 8856, eds. A. Gelbukh, F. C. Espinoza and S. N. Galicia-Haro (New York: Springer, 2015), 9–17, https://doi.org/10.1007/978−3−319−13647−9_2.

[40] Maciej Eder, Jan Rybicki and Mike Kestemont, „Stylometry with R: A Package for Computational Text Analysis," *The R Journal* 8, 1. sz. (2016): 107–121, https://doi.org/10.32614/rj−2016−−007; Maciej Eder, „Style-Markers in Authorship," 103.

[41] E témában részint a magyar nyelvre vonatkozóan is találunk megállapításokat: Jan Rybicki and Maciej Eder, „Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?" *Literary and Linguistic Computing* 26, 3. sz. (2011): 315–321, https://doi.org/10.1093/llc/fqr031.

[42] Jack Grieve. „Quantitative Authorship Attribution: An Evaluation of Techniques," *Literary and Linguistic Computing* 22, 3. sz. (2007): 251–270, https://doi.10.1093/llc/fqm020.

[43] A leggyakrabban alkalmazott elemzőmódszereket és távolságmértékeket Eder ide tartozó kutatási eredményei alapján mutatom be: Maciej Eder, „Style-Markers in Authorship," 99–114; Maciej Eder, Jan Rybicki and Mike Kestemont, „Stylometry with R: a package for computational text analysis," *R Journal* 8, 1. sz. (2016): 107–121, https://doi.org/10.32614/rj−2016−007.

 DOI 10.31400/dh-hun.2019.2.336

dezünk. Az egyes klasztereken belüli adatok valamely jellemzők mentén hasonlítanak és különböznek a többi klaszter elemeitől. Hasonló multidimenzionális eljárások a *főkomponens-analízis,* a *faktoranalízis,* a *többdimenziós skálázás,* a *diszkriminanciaanalízis,* a *Support Vector Machine (SVM),* a *Nearest Shrunken Centroids (NSC),* és Burrows attribúciós tesztjei: a *Delta, Zeta* és *Iota.*

A stilometriai elemzés során statisztikai módszerrel különféle stílusmarkerek előfordulási gyakoriságát vizsgáljuk a szövegekben, azaz a szövegekhez a stílusmarkerek terében egy-egy vektort rendelünk. Az így kapott, adott szövegre jellemző értékeket távolságmértékkel elemezzük, hogy meghatározzuk a szövegek egymáshoz való viszonyát. Két szöveg hasonlósága a tulajdonságok terében az őket reprezentáló vektorok között lévő távolsággal határozható meg.

A multidimenzionális eljárás során a szövegkorpusz gyakorisági tényezői közti távolság mérésére alkalmazott *távolságmérték* kiválasztását nagyban meghatározza az, mit akarunk elemezni. Az *euklideszi távolság* csak azokban az esetekben megfelelő, ha a markerek eloszlása a szövegekben hasonló, amely sokféle markernél nem áll fenn, pl. a szavak gyakorisága jellemzően nem ilyen. Alkalmas lehet azonban a ritka, a témát megjelölő szavakra, hiszen azok jellemzően egyenlő mértékben szerepelnek a korpuszokban. A *Manhattan távolság* már a normalizált távolságot méri. A *Classic Delta* normalizált szógyakoriságot mér, de függ az elemzett szövegek arányától és a szerzők szövegarányától. Argamon *Lineáris Deltá*ja Burrows *Deltá*jának és az *euklideszi távolság*nak a keveréke: a normalizált jelleggyakoriságokra alkalmazott *euklideszi távolság,* amely érzékeny a szövegek számára. Eder *Deltá*ja a flektáló nyelvekre jól alkalmazható, a *Classic Delta* módosítása. A *Canberra távolság* nagyon szenzitív a szerzők közti ritka szóhasználatra, és érzékeny az elemzett szavak számára.

A sikeres stilometriai elemzés közel sem triviális feladat. Nemcsak az összetett elméleti háttér alapos ismeretére van szükség, de az empirikus módon szerzett tudás is meghatározó szerepű. A megfelelő reprezentatív korpusz összeállításának a fontossága alapvető, hiszen fel kell ismernünk, hogy miből adódik a szövegek közti különbség. Ha például a vizsgált szövegek tematikailag nagyon eltérőek, akkor könnyen meglehet, hogy a témák közti különbség domborodik ki, és mégsem a szerzők közti eltérést elemeztük, ahogy terveztük. Ugyanez igaz a különféle műfajok és az időszakok közti különbözőségekre is. A vizsgálandó szövegek méretének a figyelembe vétele szintén lényeges szempont, hiszen bizonyos *távolságmérték*ek erre nagyon érzékenyek, és emiatt torzíthatnak. Az eredményességhez hozzájárul az adott vizsgálathoz legadekvátabb módszerek megtalálása és ezek kombinációja, együttes alkalmazása. Ami az egyik esetben sikeres attribúciós eljárás, az nem feltétlenül működik a másikban. Mára már megdőlni látszik az a feltételezés, hogy a stilometriai elemzés során azt a technikát kellene megtalálni, amely sikeres lehet minden műfajra, nyelvre és korszakra.[44] Helyette inkább az adott feladathoz és elemzéshez érdemes a legadekvátabb módszert kialakítani. Az ily módon elvégzett vizsgálat esetén is inkább valószínűségről, mint teljes bizonyosságról beszélhetünk, és nem nélkülözhető a kritikai szellemű nyelvi-filológiai kontroll sem.

[44] Holmes and Kardos „Who Was the Author," 5.

3.3. Mivel mérjünk?

A szerzőségi és stilometriai elemzésekhez ma már különféle informatikai eszközcsomagok állnak rendelkezésre. A bölcsészkutatók számára egyszerűen alkalmazható a magyar nyelv statisztikai alapú szövegelemzésére is alkalmas, nyílt hozzáférésű, Maciej Eder, Jan Rybicki és Mike Kestemont által R-ben kialakított *Stylo* programcsomag,[45] amelynek hazai fejlesztésben már webes alkalmazása, a *Shtylo* is elérhető.[46] Ez utóbbi előnye, hogy a futtatókörnyezet kialakításának a terhét leveszi a kutató válláról. A webes alkalmazáshoz egy böngészőre van szükség, a sok memóriát és processzoridőt igénybe vevő feladatok egy központi szervergépen futnak. További előnye, hogy a korpuszokat adatbázisban tárolja el. Ugyan az alkalmazás a munkafolyamatok különböző lépéseit elvégzi helyettünk, de a konfiguráció és a paraméterezés ezzel együtt is komoly hozzáértést és tapasztalatot igénylő feladat, amely érinti a bemenettel és a nyelvvel, a választott vizsgálandó és a leggyakoribb vizsgálandó elemekkel, a selejtezéssel, a választott statisztikai elemzőmódszerrel, a mintavételezéssel és a kimenet formátumával kapcsolatos beállításokat. A továbbiakban bemutatott kísérleteket ezzel az alkalmazással végeztük el.

4. Kísérletek a Mikes-korpuszon

A vizsgálat alapkérdése az, hogy vajon magyar történeti szövegen eredménnyel tudjuk-e alkalmazni a statisztikai elemzésnek ezt a típusát. Ehhez azt vizsgáltuk, hogy a Mikes-művek[47] különböző jellegű csoportosítása a nyelvi megformáltság alapján stilometriai eszközökkel megvalósítható-e, illetve az életművel kapcsolatban meglévő ismereteink alapján igazolható-e a módszer alkalmazhatósága.[48] A tanulmány három kísérletet tárgyal: az első a saját szerzőségű szövegek és a fordítások kapcsolatát, a második a műfaji-tematikai besoroláson alapuló beszédmód szerinti elkülönülést mutatja be. A harmadik kísérletben a stilometriai elemzések hatékonyságának a növelését vizsgáljuk, amelynek egyik lehetőségét egy olyan fázis beiktatásával képzeljük el, amelyben a digitális szótár mint történeti szöveget normalizáló eszköz jut szerephez.

A teljes Mikes-életmű betűhű kritikai kiadása mintegy 6000 oldalnyi terjedelmű és kb. 1,5 millió szót tartalmaz. A saját szerzőségű *Törökországi levelek* mellett Mikes munkásságának jelentősebb része franciából való fordításokból áll, amelyeket Hopp Lajos a következő kategóriákba sorolt: erkölcsnevelő értekező próza önálló átültetése;

[45] Eder, Rybicki and Kestemont, „Stylometry with R," 107–121, `https://doi.org/10.32614/rj--2016-007`.

[46] Az alkalmazásról részletesen: *Shtylo,* hozzáférés: 2019.02.20, `https://github.com/dobijan/shtylo/wiki`. Dobi Jan Sándor, Mészáros Tamás és Kiss Margit, „Shtylo: stilometriai elemzések webes támogatása," in *XIV. Magyar Számítógépes Nyelvészeti Konferencia,* szerk. Vincze Veronika (Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2018), 423–436.

[47] A vizsgálat a kritikai kiadás szöveganyaga alapján történt: Mikes Kelemen, *Összes művei,* s. a. r. Hopp Lajos (Budapest: Akadémiai Kiadó, 1966–1988).

[48] Az elemzések informatikai hátterét Dobi Jan Sándor „Shtylo: egy webalkalmazás az R-beli stilometriacsomag, a Stylo számára" című önálló laboratóriumi dolgozat (BME Villamosmérnöki és Informatikai Kar Méréstechnika és Információs Rendszerek Tanszék, konzulens Mészáros Tamás, 2016) taglalja.

 DOI 10.31400/dh-hun.2019.2.336

szépprózai átdolgozások; elmélkedő, didaktikus, kegyességi próza; klasszikus történeti értekező próza.[49] E műfaji-tematikai besorolást alapul véve a teljes életmű beszédmód szerinti elkülönítésben három főbb kategóriába sorolható: élőbeszéd, vallásos, erkölcsi (1. táblázat).

Műcím	Rövidítés	Tokenek száma	Saját vagy fordítás	Beszédmód
Törökországi levelek, Misszilis levelek	TL	105860	saját	élőbeszéd
Épistolák	É	268611	fordítás	vallásos
Keresztényi Gondolatok	KG	29694	fordítás	vallásos
A Kristus Jésus Életének Historiája	KJÉ	64146	fordítás	vallásos
A Keresztnek királyi uttya	KKU	160581	fordítás	vallásos
Mulattságos napok	MN	80386	fordítás	élőbeszéd
A Valoságos Keresztényeknek Tüköre	VKT	39291	fordítás	vallásos
Az Ifjak Kalauza (A, B)	IKA, IKB	182515	fordítás	erkölcsi
Catechismus Formájára valo kőzőnséges Oktatasok (A)	CA	200489	fordítás	vallásos
Catechismus Formájára valo közönséges. Oktatások (B)	CB	193533	fordítás	vallásos
Az idö Jóll el Töltésének Módgya Minden féle rendben	IJE	40872	fordítás	élőbeszéd
Az Izraéliták Szokásárol	ISZ	30333	fordítás	vallásos
A Keresztényeknek Szokásirol	KSZ	51695	fordítás	vallásos
A Sidok és az Ujj Testámentumnak Historiája	SUT	98295	fordítás	vallásos

1. táblázat. *Mikes műveinek áttekintő táblázata a korpusz mérete, a szerzőség és a beszédmód szerinti besorolás alapján*

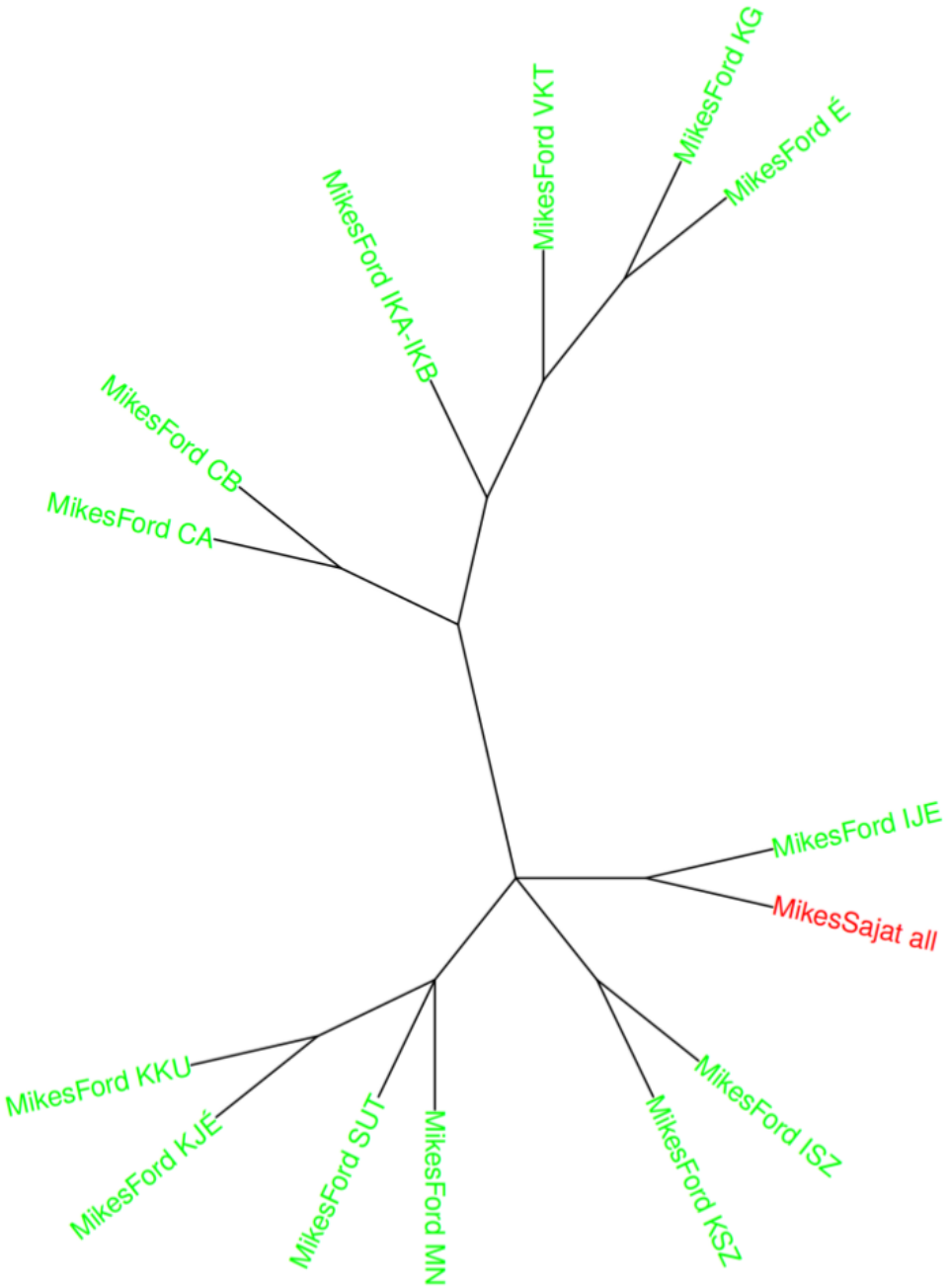
Az első kísérletben a saját szerzőségű levelek és a fordítások kapcsolatát vizsgáltuk. Mikes Kelemen fordítói munkásságát kevésbé tartják számon, holott maga az író sem határolta el egymástól alkotói tevékenysége e két területét, a levélírás és a fordítás szorosan érintkezik.[50] A szerzői életmű ugyanakkor túl nagy terjedelmű ahhoz, hogy manuális eszközökkel átfogó nyelvi vizsgálatot lehessen rajta végezni. Áttekintő elemzés elvégzéséhez számítógépes elemzőmódszerek nyújthatnak segítséget. Ennek egy korai példája az a részint számítógéppel, részint manuálisan, a szókészlet reprezentatív mennyiségén végzett lexikológiai elemzés, amely rávilágított, hogy az író saját szerzőségű munkái, a levelek és a fordítások között különbség tapasztalható a szókészlet markáns elkülönülése, az előremutató szóalkotási technikák és a szövegformálás tekintetében.[51] Jelen kutatásban arra voltunk kíváncsiak, hogy a stilometriai elemzés

[49] Hopp Lajos, *A fordító Mikes Kelemen* (Budapest: Universitas Kiadó, 2002), 133–385.

[50] Hopp, *A fordító Mikes,* 133–385.

[51] Kiss Margit, „»más értelmet adni ezeknek a szóknak«: Mikes Kelemen szóhasználatához," in *Nunquam autores, semper interpretes: A magyarországi fordításirodalom a 18. században,* szerk. Lengyel Réka (Budapest: MTA BTK Irodalomtudományi Intézet, 2016), 58–68.

hogyan tudja elkülöníteni a saját szerzőségű művet a fordításoktól, vagyis a Mikes-szókészlet teljes egészét érintő majdani vizsgálatba a stilometriai elemzés bevonható-e, s alkalmazható-e magyar nyelvű 18. századi szövegekre. Az első kísérletben a saját művek (piros) és a fordítások (zöld) (l. ábra) szétválasztására tettünk kísérletet a *Shtylo*val. Általánosságban elmondható, hogy a paraméterek beállítása egy iteratív folyamat, a beállítás helyességét az a priori tudással próbáljuk ellenőrizni.



1. ábra. *A fordítások és a saját művek elrendeződése konszenzusfán. Paraméterezés a Shtyloban: 100-800 MFW 2-grams Culled @ 0-100 %, Eder's Delta distance Consensus 0,9*

 DOI 10.31400/dh-hun.2019.2.336

Két csoportra osztottuk a műveket (1. táblázat, 1. ábra).[52] A fordításokat tartalmazó csoport jóval több művet és hosszabb szövegeket tartalmazott, mint a másik. Mivel a *Classic Delta* érzékeny a korpuszok méretére, ezért *Eder Deltá*-ját alkalmaztuk. Emellett szólt még az az érv is, hogy ez a *távolságmérték* a nem izoláló jellegű nyelvek esetében jobb eredményeket ad. Az elemzési eljárások közül a *konszenzusfá*t választottuk, amely széles körben elterjedt mód a stilometriai elemzésekben, és alkalmas arra, hogy a különböző művek közti hasonlóságot és eltérést jól láttassa. Ebben az eljárásban több egymás utáni *klaszteranalízis* fut, amelynek során több különböző beállítás mellett történik az összehasonlítás. A beállítások többségében egy adott hasonlóság kimutatható, a *konszenzusfa* ezeket ábrázolja. Ebben az elemzésben csak azokat a hasonlóságokat tartjuk meg, amelyek a beállítások többségénél megjelennek. Maga az ábrázolás nem a szövegek közti távolság nagyságát ábrázolja, hanem a hasonlóság gyakoriságát mutatja. A különböző beállításokkal végzett kísérletek minél többször mutatnak hasonlóságot, annál szorosabb kapcsolatot mutatnak, és annál közelebb helyezkednek el egymáshoz a fán.[53] Az elemzés eredményeképpen (1. ábra) a Mikes-művek négy fő csoportba különültek el. A beállítások módosításait követően is ugyanazt láttuk, hogy egyedül egy fordítás (IJE) esik közel a saját szerzőségű műhöz (TL), a futtatások 90%-a azt mutatta, hogy van köztük kapcsolat. Ez az eredmény nem hozott váratlan meglepetést abban a tekintetben, hogy a Mikes-korpusz feldolgozásával készülő *Mikes-szótár*[54] szócikkeinek írása során szoros olvasással is valószínűsíthetőnek tűnt e két mű szókészlettani közelsége, de a hasonlóság gyanújába egy másik mű is keveredett, amelyet majd az elemzéshatékonyság növelésével végzett kísérlet fog igazolni. A többi fordítás külön konszenzuságban található, ugyanakkor az látszik, hogy e művek között is fennáll a kapcsolat, amely a konszenzus erősségének beállítása során végig megmaradt, így nem rendeztük a korpuszt egymástól független művekre. Az is kiolvasható ugyanakkor, hogy a fordításvariánsok (CA, CB) – amelyek között minimális mértékű nyelvi eltérés található – egy közös ágon találhatók, ezen túlmenően a saját szerzőségű mű és a fordítások jól elkülönülnek egymástól. Az elemzés során a konszenzusküszöböt magasra állítottuk, hogy a fordítások és a saját művek közti hasonlóság a legjobban látszódjék. A konszenzusküszöb megadásánál azt határozzuk meg, hogy az elvégzett kísérletek hány százalékában jelenjen meg a hasonlóság.

Az eltérő hosszúságú szövegek (lásd az adatokat az 1. táblázatban) torzíthatják a statisztikai elemzéseket, éppen ezért az elemzés során lehetőség van a szövegek mintavételezésére, amelynek során a szöveghosszakat hasonló méretűre állítjuk be. Hogy a Mikes-szövegek eltérő hossza közti különbség ne torzítsa a statisztikai elemzést, a

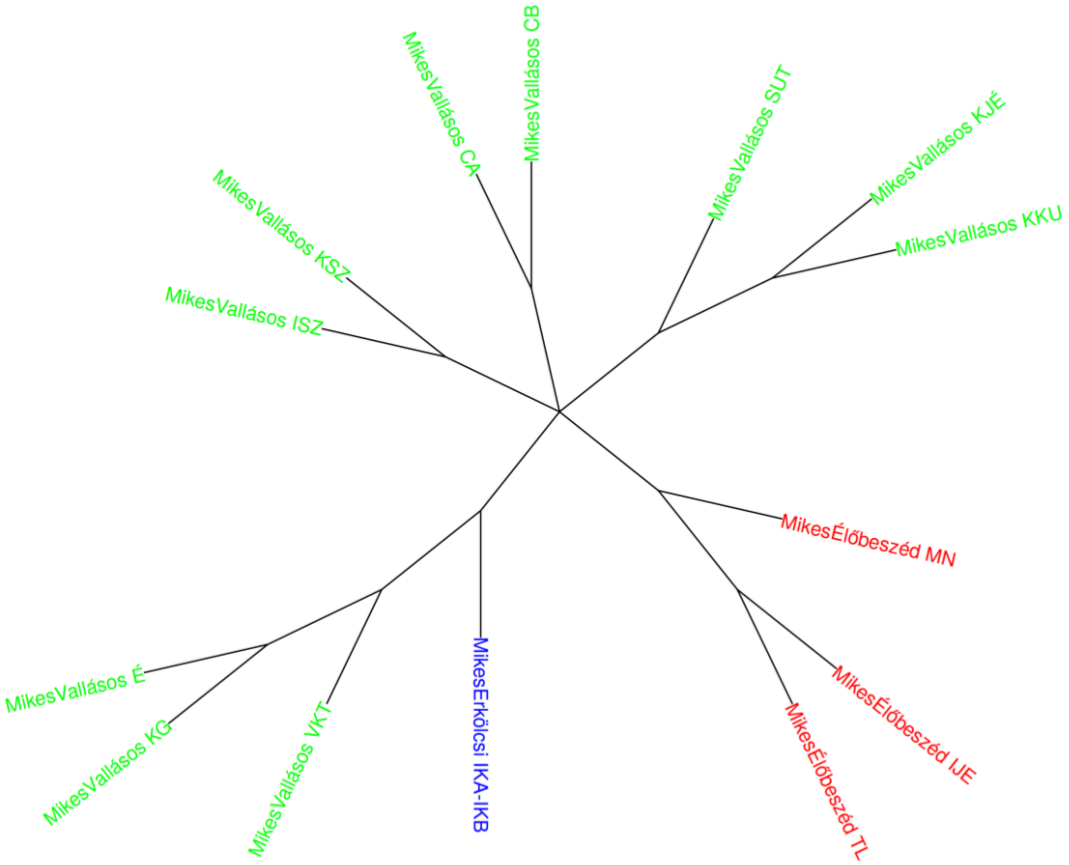
[52] A kísérletekben szereplő szövegkorpusz a mellékletben található: a digitalizált kritikai kiadás betűhű Mikes-szövegeit tartalmazza a sajtó alá rendező bejegyzései, kommentárjai nélkül.

[53] Eder, Rybicki and Kestemont, „Stylometry with R," 107–121, `https://doi.org/10.32614/rj--2016-007`; Maciej Eder, „Visualization in Stylometry: Cluster Analysis Using Networks," Digital Scholarship in the Humanities 32, 1. sz. (2017): 50–64, `https://doi.org/10.1093/llc/fqv061`.

[54] A digitális Mikes-szótár a teljes írói korpuszt feldolgozó szótár, amely 2010 óta folyamatosan készül. Jelenlegi fázisában alaki rendszerezést ad, ami azt jelenti, hogy minden mikesi szóelőfordulás mai alakú címszóhoz tartozik. Az állomány folyamatosan bővül, az eddig elkészült anyag itt érhető el: Kiss Margit szerk., *Mikes-szótár: elektronikus adatbázis* (Budapest: MTA BTK Irodalomtudományi Intézet), hozzáférés: 2019.02.20, `http://www.mikesszotar.iti.mta.hu`.

mintavételezés segítségével normalizáltuk a szöveghosszúságot (Sampling 1000). A stílusmarkerek közül a leggyakoribb bigramok (MFW 2-grams) beállítás bizonyult megfelelőnek. Biztató eredmény, hogy a stilometriai elemzés alátámasztotta a saját művek és fordítások viszonyáról meglévő eddigi ismereteinket az életművel kapcsolatban, s ez egyben azt is jelenti, hogy ezzel a módszerrel az egyes művek közti lexikai alapú hasonlóságok, különbözőségek feltérképezése további, részletes kutatás tárgyát tudja képezni a jövőben.

A következő kísérletben egy olyan vizsgálatot végeztünk, amelyben a mikesi életmű darabjain a beszédmód szerinti elkülönülést kívántuk láttatni (1. táblázat, 2. ábra). Arra voltunk kíváncsiak, hogy a *Shtylo* segítségével lehetőségünk van-e az író életművében jól elkülöníthető egyházi, erkölcsi tematikájú és élőbeszédszerű műveket a szókészlet elkülönülése alapján statisztikai szempontból is igazolhatóan megkülönböztetni.



2. ábra. *A művek tematikus elrendeződése konszenzusfán. Paraméterezés a Shtyloban: 100-1000 MFW 2-grams, Culled @ 0-80%, Canberra distance, Consensus 0,5*

A Mikes-korpuszt három élőbeszédszerű (piros), egy erkölcsi (kék) és tíz vallásos mű (zöld) alkotja. Mivel a tematikai meghatározottság ebben az esetben erősen a tartalmas szavak vizsgálatára helyezi a hangsúlyt, így a *Canberra távolság* tűnt a legadekvátabbnak. A három csoport ez esetben ugyancsak eltérő hosszúságú műveket tartalmazott (1. táblázat), így ezt mintavételezéssel kompenzáltuk (lásd az előző kísérletben leírtakat), hogy a statisztikai elemzés ne torzuljon. Az eredmény vizualizálására itt is a *konszenzusfa* tűnt megfelelőnek. A konszenzusküszöb értéke ebben a kísérletben

 DOI 10.31400/dh-hun.2019.2.336

alacsony (0,5), mert az volt a kérdés , hogy a *Canberra távolság* alkalmazásával a művek kapcsolatban lesznek-e egymással, vagy távol kerülnek. Az látszik, hogy a *Shtylo* segítségével az élőbeszédszerű szövegeket (MN, IJE, TL) jól külön tudtuk választani a többitől, továbbá az erkölcsi témájú szöveg (IKA, IKB) egy ágba sorolódik a vallási témájú szövegek egy részével (É, KG, VKT), ami igazolhatja azt is, hogy ez a fajta tematikai megkülönböztetés nem jár feltétlenül a szókészlet jelentős elkülönülésével. Ebben a kísérletben a stilometriai elemzés arra volt képes, hogy az élőbeszédszerű szövegeket markánsan elkülönítse a többitől, s ez tekinthető a legerősebb stilisztikai markernek ebben a kísérleti korpuszban. Ez esetben az élőbeszédre jellemző csoportban a saját szerzőségű művek mellett ott találunk két fordítást is.

A stilometria történeti fejlődésében fontos szerep jutott a mennyiségileg meghatározható jelenségek, a szerzői megkülönböztető jegyek meghatározásának – állítja Holmes –, s ebben a tekintetben a lexikális jegyek túlsúlyba kerültek, ám az utóbbi időszakban a szintaktikai, szemantikai, grammatikai, szófajtani, morfológiai megkülönböztető jegyek is megjelentek, amelyek elemzéséhez egyre több informatikai támogatás kínálkozik, és amelynek eredményeképpen az összetett elemzések pontosabb, megbízhatóbb eredmények elérését teszik lehetővé.[55] Minthogy a gépi feldolgozásra alkalmas szövegek mennyisége folyamatosan nő, fejlődik, és egyre hatékonyabbá válik a stilometriai módszerek eszköztára,[56] így lehetővé válik a szerzőségi vizsgálatok elvégzése nagyméretű szövegkorpuszokon is.[57] A lexikai alapú elemzés javításának egyik lehetséges, további módját a harmadik kísérlet mutatja be. Ha a statisztikai szövegelemzés szógyakoriság-alapú vizsgálatai során az adott szövegkorpuszban megjelenő szóelőfordulásokkal számolunk, akkor történeti szövegek esetében különösen nagy alaki változatossággal találkozunk, Mikes estében például *ekepen, eképen, e képen, ekeppen, eképpen, e képpen, ekkepen, ekképen, ekképpen*. Ha ezt a sokféleséget a történeti alakok normalizálásával csökkenteni lehetne, akkor az elemzés hatékonyságát növelhetnénk, mivel a szóalakok változatainak a redukálásával csak az alapalakban álló szavakat (pl. *ekképpen*) hasonlítanánk össze és nem az alakváltozataikat, illetve paradigmatikus alakjaikat is (pl. *ekepen, eképen, e képen, ekeppen, eképpen, e képpen, ekkepen, ekképen*), mintha külön szótári alakok lennének. Ezt az előfeldolgozást támogathatják a gépi morfológiai elemzők is, ám magyar történeti szövegek esetében az ilyen jellegű automatizált elemzés közel sem egyszerű megoldás.[58] A történeti szövegek gépi automatikus morfológiai elemzését más, megbízhatóbb megoldással is pótolhatjuk, például ha a normalizálást szótár segítségével végezzük el. A Mikes-korpusz normalizálásához a készülő digitális *Mikes-szótár* segítséget ad, hiszen alaki rendszerezés révén minden egyes szövegben szereplő régies alakú szót mai címszóhoz rendel, ezáltal a stilometriai elemzés során nem a régies, paradigmatikus alakban

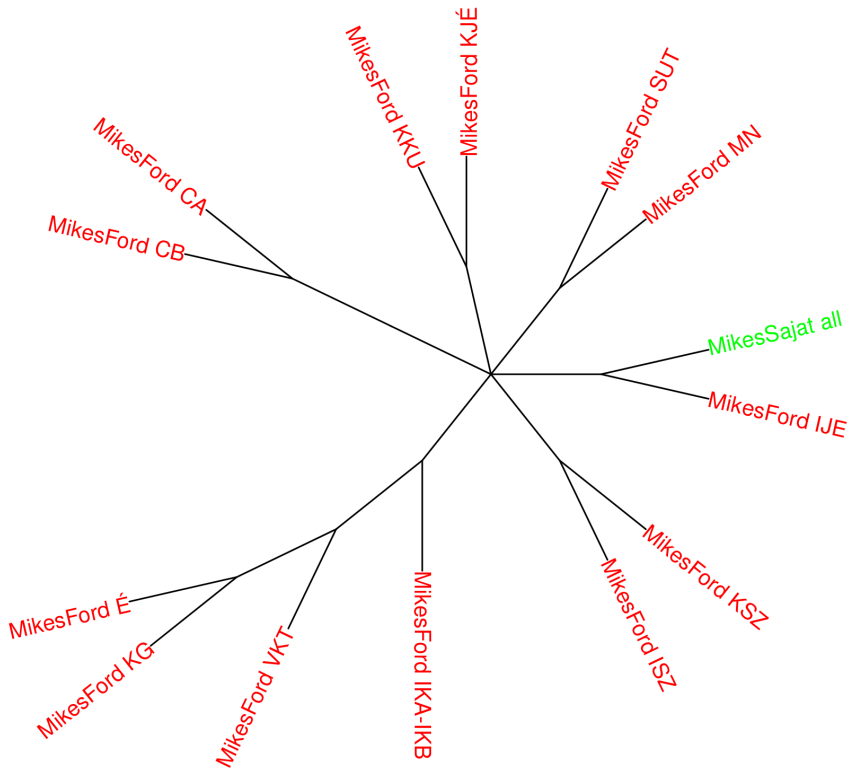
[55] David Holmes, „Authorship Attribution," *Computers and the Humanities* 28, 2. sz. (1994): 87–106, https://doi.org/10.1007/bf01830689.

[56] Évről évre újabbak látnak napvilágot, pl. Justin Stover and Mike Kestemont, „The Authorship of the Historia Augusta: Two New Computational Studies," *Bulletin of the Institute of Classical Studies* 59, 2. sz. (2016): 140–157, http://dx.doi.org/10.1111/j.2041–5370.2016.12043.x.

[57] Craig, „Stylistic Analysis," 280; MacDonald Pairman Jackson, „Determining Authorship: A New Technique," *Research Opportunities in Renaissance Drama* 41 (2002): 1–14.

[58] Kiss Gabriella, Kiss Margit és Pajzs Júlia, „A Nagyszótár történeti korpuszának elemzéséről," *Magyar Nyelv* 100, 2. sz. (2004): 185–191.

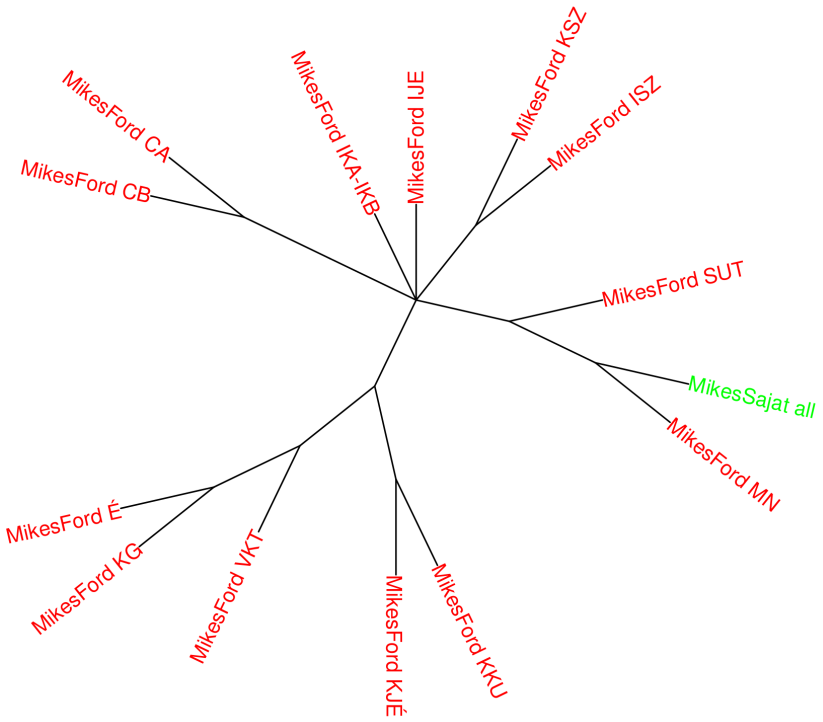
lévő szavakat hasonlíthatjuk össze egymással, hanem a mai szótári alapalakokat. Így megbízhatóbb eredményt kaphatunk az életmű szókészletére alapuló vizsgálattal kapcsolatban. Az utolsó kísérletben tehát a stilometriai elemzést kiegészítettük egy olyan előzetes munkafázissal, amelyben a *Mikes-szótár* segítségével végeztük el a szövegek normalizálását.[59] A szavak szövegbeli előfordulási alakjait helyettesítettük a standardizált, mai szótári alakban (nem toldalékolt!) álló megfelelőikkel annak érdekében, hogy növeljük a statisztikai elemzés hatékonyságát a szavak közti különbségek redukálásával, amely a toldalékolás és a történeti szöveg egyenetlensége miatt van jelen. Ebben a kísérletben két elemzést végeztünk ugyanazokkal a beállításokkal, hogy összehasonlíthatóvá váljék a különbség a két futtatás között.



3. ábra. *Mikes-művek elemzése az eredeti szövegek felhasználásával. Paraméterezés a Shtyloban: 100-2000 MFW 2-grams, Culled @ 0%, Classic Delta distance, Consensus 0,5*

[59] Margit Kiss and Tamás Mészáros, „Creating an Extended Author's Dictionary to Support Digital Literary Research," *Abstracts of DH Benelux Conference, June 6–8, 2016,* hozzáférés: 2019.02.20, `http://2016.dhbenelux.org/wp-content/uploads/sites/4/2016/05/89_Kiss-Meszaros_FinalAbstract_DHBenelux_2016_long.pdf`.

 DOI 10.31400/dh-hun.2019.2.336



4. ábra. *Mikes-művek elemzése a normalizált szóalakokkal. Paraméterezés a Shtyloban: 100-2000 MFW 2-grams, Culled @ 0%, Classic Delta distance, Consensus 0,5*

Az első esetben az eredeti Mikes-szövegeket elemeztük (3. ábra), a másodikban a szótári szavakra lecserélt normalizált változatot (4. ábra). A leggyakoribb vizsgálandó elemeknél szintén a leggyakoribb bigramok (MFW 2-grams) beállítást választottuk. A *Classic Delta* távolságmértéket alkalmaztuk, amelyet a normalizálás indokolttá tesz. A szövegkorpusz egyenetlenségét mintavételezéssel kompenzáltuk. Az elemzés eredményének a vizualizálásához a konszenzusfát alkalmaztuk. A konszenzusküszöb kialakításánál arra törekedtünk, hogy az életmű egyes darabjai kapcsolatban maradjanak egymással. Az elemzés eredményéből látható, hogy mindkét futtatásnál a CA, CB fordításvariánsok értelemszerűen nagyon közel maradt egymáshoz, amely az elemzés relevanciáját erősíti, hiszen nagyon minimális eltérés van a két szöveg között. Az eredeti szövegek elemzésénél az É, a KG, VKT ágához az IKA, IKB variánsai esnek közelebb. Míg a normalizált korpuszon az IKA, IKB variánsai helyet cserélnek a KJÉ és a KKU írásokkal. További látványos különbség, hogy a saját szerzőségű művek (TL) az eredeti szövegeket tartalmazó korpuszvizsgálat esetében IJE írással vannak legközelebbi kapcsolatban (ahogy a saját művek és a fordítások esetén is láthattuk), ugyanakkor a normalizált korpuszon végzett elemzés során az MN esik hozzá legközelebb, továbbá ugyanazon az ágon található még kicsit távolabb a SUT. Az eredeti szövegvizsgálat során a SUT és az MN került szoros kapcsolatba egymással.

A szótár szerkesztése során empirikus megfigyeléssel is érzékelhető volt, hogy a *Törökországi levelek* (TL) lexikális anyaga szorosabb kapcsolatban van azokkal a fordí-

tásokkal, amelyeket az itt bemutatott stilometriai elemzések eredményeztek. Mindennek további, mélyreható és átfogó feltárásához megvan a kiindulási eszköz, amely a terjedelmes életmű módszeres feldolgozásának egyik lehetséges módja. Az itt bemutatott példák a stilometriai módszerek Mikes-korpuszra történő alkalmazhatóságát támasztják alá. Jelen keretek között nem cél a bemutatott eredmények további, mikesi életművel kapcsolatos mikrofilológiai elemzése, ugyanakkor egy nyelvi-alkotói folyamatokat feltáró későbbi, további adatelemzésen alapuló munka kezdeti lépéseként értelmezendő. Az írói munkásságot terjedelme miatt nyelvi szempontból részleteiben, egyes aspektusaiból vizsgálták ez idáig.[60] E munkának további kutatási iránya lehet a stílus fogalmának, értelmezési kereteinek a továbbgondolása, amely az új módszertannak köszönhetően is formálódik.[61] A digitális korpusz és a szótár segítségével, valamint az informatikai támogatású elemzőmódszerek alkalmazásával lehetőség nyílik nagyobb léptékű vizsgálatok elvégzésére a jövőben. Egyúttal ez azt is jelenti, hogy a digitális szótárakkal szembeni elvárásokat, feladatokat is revideálnunk kell. A digitális szótár ellátja a hagyományos szótári funkciókat, ezen túlmenően strukturált szövegkorpuszként az informatikai alapú szöveg- és korpuszelemzést, annak hatékonyságát növelő eszközként is képes támogatni.[62]

5. Összegzés

A szépirodalmi szövegeket feldolgozó stilometriai kutatások gyakran heves viták kereszttüzébe kerülnek, sokan a létjogosultságukat is kétségbe vonják, holott ha segédeszközként tekintünk rájuk a filológiai vizsgálatokban, és nem egyedüli módszerként, akkor árnyaltabb képet kaphatunk e területről.[63] A dolgozatnak ezeket az anomáliákat nem volt célja bemutatni, helyette inkább azokra az eredményekre koncentrált, amelyek azt támasztják alá, hogy a statisztikai alapú szerzőségi, stilometriai elemzés olyan eljárások közé tartozik, amely támogatni képes a szoros olvasás során vizsgálandó problémák megoldását. Ennek érdekében e tudományterület jelenlegi eredményeinek és a kísérletek hátterének módszertani feltárásához nyújtott áttekintésén túl konkrét stilometriai elemzéseket is bemutatott a dolgozat, amelyek alátámasztották az elvégzett kísérletekben e módszer relevanciáját. A továbblépés egyik lehetséges iránya az, hogy még pontosabbá tegyük a stilometriai elemzést, amelynek például

[60] Például Szabó T. Attila, „A székely nyelvjárások a magyar irodalomban," *Új Látóhatár* 4 (1989): 549–557.

[61] Nemzetközi diskurzusban pl. Berenike Herrmann, Karina van Dalen-Oskam and Christof Schöch, „Revisiting Style, a Key Concept in Literary Studies," *Journal of Literary Theory* 9, 1. sz. (2015): 25–52, https://doi.org/10.1515/jlt−2015−0003.

[62] Margit Kiss and Tamás Mészáros, „Rethinking the Role of Digital Author's Dictionaries in Humanities Research" in *Proceedings of the XVIII EURALEX International Congress,* Simon Krek, Jaka Čibej, Vojko Gorjanc and Iztok Kosem eds. (Ljubjana: Ljubjana University Press, 2019), 871–880; Mark Andrew Algee-Hewitt, „The Hidden Dictionary: Text Mining Eighteenth-Century Knowledge Networks" in *Digital Humanities 2018, DH 2018, Book of Abstracts,* eds. Jonathan Girón Palau and Isabel Galina Russell (Mexico City: El Colegio de México, UNAM, and RedHD, 2018), 146–147.

[63] A többféle elemzés kombinációján alapuló vizsgálatot és a vizsgálati eredmények valószínűségéről ír Patrick Juola, „The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions," *Digital Scholarship in the Humanities* 30, 1. sz. (2015): 100–113, https://doi.org/10.1093/llc/fqv040; Grieve, „Quantitative Authorship Attribution," 251–270, https://doi.10.1093/llc/fqm020.

 DOI 10.31400/dh-hun.2019.2.336

egyik módja lehet a digitális szótár bevonása a történeti szöveg normalizálásába. Másik lehetséges irány a kapott eredmények felhasználásával további nyelvi-filológiai elemzések elvégzése, amellyel a hatalmas életművel kapcsolatos tudásunk tovább gazdagodik, ez azonban már e dolgozat keretein túlmutató feladat.

> Sokat vitatott dolog még mindig, hogy azok, akik az irodalommal foglalkoznak, miért nem szeretik a számok világát: nem bíznak bennük, és nem értik, hogy az ember hogy tud valami olyan banalitással foglalkozni, mint hogy irodalmi szövegekben megszámoljon valamit.[64]

The Potentials of Stylometry Analysis of Hungarian Historical Text Corpora

This paper discusses the potentials of the computer assisted analysis of Hungarian historical texts, which relies on the methods of linguistic, literary studies, information technology and statistics. It reviews the characteristics and applicability of different stylometric methods and demonstrates their use by case studies based on Kelemen Mikes' (1690-1761) works. It examines the relationships between Mikes' own works and his translations as well as the thematic separation of his works. The case studies highlight that the effectiveness of the stylometric analyses of Mikes' writings can be improved by applying the digital Mikes' dictionary.

Keywords:
authorship attribution, stylometry, digital author's dictionary, Kelemen Mikes

[64] Hugh Craig, a University of Newcastle professzor emeritusa, egyetemi weboldalán *Figures of speech* címmel megjelent összegzés, hozzáférés: 2019.02.20, `https://www.newcastle.edu.au/profile/hugh-craig`, (ford. tőlem).

Tóth Tünde

Hankuk Egyetem, Dél-Korea

tunde.toth@hotmail.com

Életünk a Kínai Szobában[*] I. Odi et amo

Míg a múlt század végén az információtechnológiai kutatások lényegében optimista, építő munkát jelentettek, és az úttörő kutatók számára egy alapvetően pozitív cyberjövő képe sejlett fel, ahol a technológia elsősorban az információ- és tudáshozzáférést segíti, addig a 21. század embere számára Searle „Kínai Szoba" argumentuma lassan oly módon lesz a mindennapok valósága, hogy a bennünket körülvevő cybervilágban elveszítjük a kontrollt a bemenet és/vagy a kimenet vagy az információ, a tartalom, sőt, néha éppen a saját privát szféránk felett.

Jelen tanulmány első fele a gyűlölet különböző formáinak online térben való megjelenését elemzi. A szerző a koreai popzene világában történt „Jung Joon-yung-botrány" esetén keresztül tárgyalja a virtuális megszégyenítés *(online shaming)* néhány formáját a virtuális kukkolástól *(cyber-stalking)* a titokban készített pornográf felvételek *(nonconsensual pornography)* megosztásán át a nyílt online gyalázkodásig *(cyber-pillory)*. A dolgozat második része a baudrillard-i szimulakrum-fogalom segítségével a virtuális szerelem egyes megnyilvánulási formáit elemzi a koreai sztárok elvadult rajongóinak *(sasaeng)* esetétől a robotok és más képzetes hősök iránt érzett szerelemig. A virtuális valóság nem pusztán egyszerűbbé teszi a különböző emberi érzelmek kifejezését azáltal, hogy a valós kommunikációt körülményessé tevő tényezők egy részétől, például az önreflexiótól megszabadulunk, ám egyúttal le is egyszerűsíti mondanivalónkat („nem tetszik" – „imádom"), s ezzel tulajdonképpen magára a gondolkodásunkra is visszahat. Más emberré válunk azáltal, hogy olyan dolgokat is megteszünk a virtuális térben, amit a valóságban nem tennénk. A szimulakrum problémája a „Nem vagyok robot", az „Ember vagy te is?" és a „W" című koreai sorozatokon keresztül kerül bemutatásra, s az elemzés végén *az emberi tudat digitálisan örökéletűvé tételének* etikai problémáiról is szó esik.

Kulcsszavak:
digitális privát szféra, közösségi háló, szimulakrum, K-pop, K-dráma

[*] Esther Dyson könyvének volt *Életünk a digitális korban* az alcíme, melyet azzal reklámoztak, hogy „nehezen meghatározható műfajú könyv" egy amerikai „Internet-guru" tollából. Címadásom tisztelgés akar lenni azok előtt, akik már akkor az „információs szupersztrádán" jártak, amikor az még csak egy poros, szűk kis ösvény volt a tartalomínség pusztájában. Esther Dyson, *2.0 verzió: Életünk a digitális korban,* ford. Kozma Zsolt (Budapest: HVG, 1998).

Searle 1980-as Kínai Szoba-argumentuma[1] szellemes érv a Turing-teszt, vagyis Turing azon elgondolása[2] ellen, hogy „a diszkrét állapotú szimbolikus jelfeldolgozást végző számítógépek kognitív állapotokkal rendelkezhetnek."[3] Nagyon leegyszerűsítve, Turing szerint *ha úgy tűnik nekem, hogy ember, akkor ember; ha úgy tűnik nekem, hogy gondolkodik, akkor gondolkodik* – ami voltaképpen Descartes *cogito*jának (1637) egyfajta sajátos parafrázisa. A fenti állítások mindegyikét kétségbe vonták eddig: Turing feltevése egy *entiméma*, melynek egyik burkolt előfeltevése, hogy az ember mint külső szemlélő képes eldönteni, hogy géppel vagy emberrel áll-e szemben; a másik pedig az, hogy az ember gondolkodó lény. Nem arról van szó, hogy néha úgy tűnik, nem az, hanem egyszerűen – minden törekvésünk dacára – a legtöbb esetben nem rendelkezünk a) pontos, b) teljes, c) szükséges vagy elégséges információval. Nagyon kevés tudományágról mondható el, hogy következtetései (megoldásai) teljes és pontos információkon alapulnak. A klasszikus bölcsészettudományok nagy részében hol a tárgy, hol pedig az információk egészének hozzáférhetetlensége folytán kénytelenek vagyunk közelítésekkel és hipotézisekkel dolgozni. Ugyanakkor a tárgy és az információ jellegéből adódóan számos esetben nem az egzakt leírás (pl. a jelenséget leíró képlet), hanem a *kreatív, de bizonyítható* interpretáció a tevékenység célja.

A világhálón történő kommunikáció (beleértve az okostelefonosat is) számos esetben új kihívások elé állít bennünket. Sokszor persze *a való életben* (bevett rövidítéssel élve: IRL, azaz „in real life") is nehéz eldönteni, mi a másik ember célja, mik a szándékai, ám többnyire abból a feltételezésből indulunk ki, hogy még a legelemibb kommunikációs helyzetben is léteznek a másiknak céljai. A gépeken keresztül történő *személyes* vagy *tömeg*kommunikáció esetében ezek a szándékok kevésbé megfejthetők, és jobban elleplezhetők.

Az ingyenes fordítóprogramokkal készített adathalász leveleket egyelőre azonnal elárulja a levelek stílusa, még ha nem is minden címzett képes arra, hogy a nyelvtani és stiláris hibákat észrevegye. (Érdekes módon ugyanez mondható el a hivatalos nyelvezetet rosszul használó, anyanyelvükön rosszul fogalmazó adathalász bűnözőkről is.)

Találkozhatunk azonban olyan professzionális, okostelefonra készült kommunikációs alkalmazásokkal is, melyek emberi közreműködéssel, szakfordítók által összeállított készletből dolgozva adnak kész, mind grammatikailag, mind stilárisan helyes mondatokat. Ezek valójában inkább elektronikus társalgási zsebkönyvek, mintsem fordítóprogramok. Akinek először volt része ilyen gépi segítséggel történő kétirányú kommunikációban, az bizonyára sokáig emlékezni fog arra a szürreális élményre, amelyet ezek a programok nyújtani képesek az egymás nyelvét, kérdéseit, válaszait, sokszor még gesztusait sem értő felhasználóknak.

Még közelebb visz bennünket a Kínai Szobához a különböző kommunikációs tanuló algoritmusok (mint pl. az *Eviebot*) esete. Ezek a korábban más felhasználók által bevitt

[1] Mivel a példát gyakorta idézik, úgy vélem, helyes nagy kezdőbetűkkel megkülönböztetni az összes más, lehetséges „kínai szobá"-tól. John Searle, „Minds, Brains and Programs," *Behavioral and Brain Sciences* 3, 3. sz. (1980) 417–457, https://doi.org/10.1017/S0140525X00005756.

[2] Alan Turing, „Computing Machinery and Intelligence," *Mind* 59, 236. sz. (1950), 433–460, https://doi.org/10.1093/mind/LIX.236.433.

[3] Csáji Balázs Csanád, *A mesterséges intelligencia filozófiai problémái* (Budapest: ELTE BTK, témavezető: Farkas Katalin, 2002), http://old.sztaki.hu/~csaji/CsBCs_MI.pdf,12.

 DOI 10.31400/dh-hun.2019.2.387

mondatokat (kérdéseket és válaszokat) alkalmazzák a későbbi felhasználókkal történő kommunikációban, így még emberibbnek tűnnek válaszaik és kérdéseik, mint az elektronikus társalgási zsebkönyvek sablonos mondatai. Emiatt akár meg is inoghatunk, és már-már elkezdünk ezeknek a csevegőbotoknak öntudatot és szándékot tulajdonítani. Nem azért, mert az ténylegesen létezne, hanem mert mi, emberek ilyenek vagyunk. Ahogy Tomasello mondja: az ember szándéktulajdonító lény.[4]

A Kínai Szoba, azaz *a tudat a gépben* témáját egy hosszabb tanulmányban, különféle megközelítéseken keresztül kívánom végigjárni, s ennek során arra keresem a választ, minek is tekintjük a rohamléptekkel fejlődő mesterséges intelligenciát. Nem pusztán az a kérdés, hogy egyszerűen eszköznek vagy éppen önálló, tudattal bíró személyiségnek látjuk-e, vagy szeretnénk látni az AI-t *(artificial intelligence),* hanem az, hogyan viszonyulunk hozzá mi, emberek, illetve hogyan változtatja meg a gép az életünket, s egyúttal vele minket, magunkat. Míg a '90-es években sokan elsősorban arra koncentráltunk, hogy a világháló *(internet)* segítségével milyen változás következik be a kulturális emlékezetben (és a jelenséget nem is a nyomtatás, hanem az írás feltalálásának forradalmához hasonlítottuk), mára a különböző gépi technológiák segítségével elkövethető bűncselekmények, illetve a digitális térben bekövetkező személyiségi jogi sérelmek és kockázatok kerültek előtérbe. Erre a kérdéskörre egyelőre gyakrabban használjuk az angol *digital privacy* kifejezést, mint a magyar „digitális privátszféra" megfogalmazást. A régi írás-hasonlathoz visszatérve, bizonyos tekintetben a kés feltalálásához hasonlíthatjuk a digitális technológia térnyerését, s ez a „kés 2.0" rengeteg feladatot ad a jogtudománynak és a bölcsészettudománynak is.

Az elmúlt negyedszázad során létrejött digitális bölcsészet[5] *(digital humanities)* nevű humanióra mibenlétének és mineműségének definíciói kezdettől igyekeznek elhatárolódni attól, hogy a bölcsészettudományban pusztán publikációs eszközként jelenjék meg a számítógép és a világháló. A '90-es években Horváth Iván és az egykori BIÖP kutatóinak köre a néhai Neumann-ház és a saját szolgáltatásának jellegét évtizedek óta őrző Magyar Elektronikus Könyvtár (MEK) azon kiadásait, melyek online tették elérhetővé valamely szerző korábban papírformában megjelent kritikai kiadását, nem tekintették ún. *online kritikai kiadás*nak. Bizonyos szempontból[6] azt mondhatjuk, hogy az elektronikus, vagy digitális feldolgozottság, a külső és belső hiperhivatkozásokkal való teleszőttség *mérték*ét tekintették fokmérőnek, nem pedig a *gépi felületen történő távoli elérés* jóval kisebb befektetéssel produkálható kritériumának megvalósulását.[7]

[4] Michael Tomasello, *Gondolkodás és kultúra,* ford. Gervain Judit (Budapest: Osiris, 2002).

[5] 1997-ben az ELTE BTK BIÖP (Bölcsészettudományi Informatika Önálló Program) körében még *bölcsészettudományi informatiká*nak neveztük, 2005-ben pedig (a terület egy részét) *informatikai irodalomtudomány*nak (vö. MTA I.O. Informatikai Irodalomtudományi Munkabizottság).

[6] Ezenkívül természetesen a legnagyobb különbség abban rejlett, hogy a BIÖP kiadványaiban új edíciók készültek, a szakemberek filológiai munkát végeztek, míg a MEK és a Neumann-ház csak online hozzáférést biztosított papírkiadásban már meglévő anyagokhoz. Textológiai szempontból a BIÖP kiadványai javítottak a szövegeken, a többiekéi gyakran szövegromlásokhoz vezettek. (Lásd *Filológia és digitális barbárság* (2004), `http://magyar-irodalom.elte.hu/biop/barbar/`.) Emellett az online közzétett könyvek struktúrája a nyomtatottét másolta, és a szöveg dimenzióját nem terjesztették ki a hálózat mint közeg által lehetővé tett irányokba.

[7] Horváth Iván, *Magyarok Bábelben* (Szeged: JATE Press, 2000); Horváth Iván, *Gépeskönyv* (Budapest: Balassi Kiadó, 2006); Tóth Tünde, „Online kritikai szövegkiadás Magyarországon az ezredfordulón,"

A számítógép napjainkban mégis alapvetően mint eszköz van jelen a digitális bölcsészettudományban: maga a gép nem végez kutatásokat és nem jut önálló eredményekre. A Turing-teszten tehát csúfosan elbukna, viszont jóval több segítségünkre lehet, és jóval több feladatot adhat annál, minthogy a) papírkönyv helyett képernyőről olvashatunk, b) nem kell bemenni a könyvtárba, c) a zsebünkben vihetjük el a világ túlsó felére a teljes magyar kultúrkincs legjavát. Negyedszázad elmúltával egyre többen értik ezt a világháló kora előtt született bölcsészek közül is.

A számítógépes filológia,[8] a számítógépes nyelvészet[9] a világháló korának ismeretfilozófiai kutatása,[10] illetve a megismeréstudományi kutatások[11] voltak az 1990-es évek vezető digitálisbölcsészet-irányzatai[12] Magyarországon,[13] s gyakorlatilag az elektronikus könyvtárak építése és az ezekkel kapcsolatos elméleti megfontolások álltak a 20. század végén a könyvtártudomány érdeklődésének homlokterében.[14] Mára önálló részterületté vált a világon mindenütt folyó digitális bölcsészeti kutatások feltárása és ismertetése,[15] és magának az internetnek mint médiumnak a bemutatása, illetve értelmezése.[16]

Korunkra nyilvánvalóvá vált az is, hogy a *gép* és a *hálózat* olyannyira mindennapi életünk „szerves" része lett, hogy nem maradhatunk csupán a klasszikus, akadémikus tudományok talaján, ha a *gép* és az *ember* viszonyáról szeretnénk beszélni.

* * *

Helikon 5, 3. sz. (2004), 417–441; Parádi Andrea, „Internetes kritikai kiadások," `http://www.tankonyvtar.hu/hu/tartalom/tkt/magyar-irodalom/ch13s05.html`; Maróthy Szilvia, „Elektronikus szövegkiadások a könyvtárban," *Tudományos és Műszaki Tájékoztatás* 64, 6. sz. (2017), 298–309; Golden Dániel „Digitális bölcseleti értekezés," in */HI70/Tanítványok: Tanulmányok Horváth Iván 70. születésnapjára,* szerk. Bartók Zsófia Ágnes, Bognár Péter, Maróthy Szilvia (Budapest: Q. E. D. Kiadó, 2018), hozzáférés: 2019.05.15, `http://hi70.hu/2018/03/21/golden/`.

8 Horváth Iván Szegedi Számítógépes Munkacsoportja (JATE) az 1970-es évektől az 1990-es évek elejéig működött, majd az ELTÉ-n az 1990-es évek elejétől induló, 1997-ben oktatási programként is megjelenő BIÖP említhető, mely a 2000-es évek közepéig működött.

9 Prószéky Gábor és munkatársai, MorphoLogic, ill. PPKE Információs Technológiai Kar.

10 Nyíri Kristóf és munkatársai, ELTE, ill. MTA Filozófiai Kutatóintézet.

11 Pléh Csaba és munkatársai, ELTE, SZTE.

12 A nem digitális szférában is nehéz definiálni a bölcsészet- és a társadalomtudományok határait. Számos tudományágat, pl. a nyelvészetet bizonyos szempontok alapján az egyik, más szempontok alapján a másik kategóriába sorolják. Ezért fenti felsorolásom természetesen vitatható lehet. Más kérdés, hogy mind a társadalom-, mind a bölcsészettudományokat humánióráknak tekinthetjük.

13 Elnézést kérek mindazoktól, akiknek a nevét, munkásságát hely hiányában nem említettem. Mivel számos olyan műhelyben folynak DB-kutatások, melyek gyökerei korábbra nyúlnak vissza, önálló művet kívánna ezek kimerítő ismertetése.

14 Fodor János, *Trendek és tendenciák, kialakult modellek és lehetséges stratégiák az internetes közművelődési tájékoztatásban* (Budapest: ELTE BTK Irodalomtudományi Doktori Iskola, Könyvtártudományi Program, 2005).

15 김현 & 임영상, 디지털 인문학 입문 = Digital humanities [Kim Hyun és Yim Young Sang, „Bevezetés a digitális bölcsészetbe = Digital humanities"], in *Digital humanities* (Szöul: Huebooks, Hankuk Egyetem, 2016), 510.

16 Szűts Zoltán, *A világháló metaforái: Bevezetés az új média művészetébe* (Budapest: Osiris, 2013); Szűts Zoltán, *Online: Az internetes kommunikáció története, elmélete és jelenségei* (Budapest: Wolters Kluwer, 2018).

 DOI 10.31400/dh-hun.2019.2.387

Jelen dolgozatban olyan témákkal foglalkozom, amelyek általában megrekednek az újsághírek, a bulvárlapok és közösségi médiában terjedő moralizáló mémek szintjén. Az „ismeretszerzés felgyorsult tempójához és egyre szélesedő horizontjához”[17] igazodva, s természetesen a téma jellegénél fogva nem kerülhetem meg, hogy forrásaim között nem tudományos munkák is legyenek, illetve a probléma megértéséhez bizonyos mértékig le kell merülni a bulvármédia és a rajongói *mainstream* kultúra világába.[18]

Gyűlölet, 복수[19]

„Eltitkolni annyi, mint úgy tenni, mintha nem lenne az, amink van.”
(Jean Baudrillard: *A szimulákrum elsőbbsége*)[20]

A pszichológiától a jogtudományig különböző bölcsészet- és társadalomtudományok igyekeznek lépést tartani a technikai fejlődésnek azzal az árnyoldalával, hogy azonnali elérést, közvetlen hozzáférést nyújt jóhoz és rosszhoz egyaránt. Nemcsak a *nagy crux*[21] kérdéseit lehet a világháló segítségével kutatni, a bűn is beférkőzött a szobánkba, ott van a kezünk ügyében.

Az elmúlt évszázadok művészetében a fény és az árnyék együttes jelenlétével oly sokszor nyomasztónak ábrázolt nagyvárosi életforma[22] a digitális lefedettségnek köszönhetően mindenütt velünk van, és a saját otthonunkban sem rejtőzhetünk el a gonoszság elől. Sokkal közvetlenebb és azonnali hozzáférésünk, pontosabban: kitettségünk lett gyűlölőink *(haters)* és zaklatóink *(bullies)* bennünket érő verbális és virtuális terrorjának, a kukkolók *(voyeurs, stalkers)* pedig bármikor figyelhetnek bennünket. Nemcsak a *Nagy Testvér*től, vagy a technológiájuk tökéletesítésére minden felhasználójukat gépi elemzések céljára lehallgató nagy cégektől, hanem magánemberektől is féltenünk kell *digital privacy*-nket.

A hagyományos fogyasztói társadalomban a fogyasztó elvesztette a kontrollt a fogyasztási javak előállítása felett. Nem tudjuk, hogy tényleg azt kapjuk-e, amit szeretnénk, és nem tudjuk, hogy az adott termék előállítása számunkra elfogadható módon történt-e. Nap mint nap kiderül, hogy egy adott termék előállítása és/vagy szállítása mérgező, környezetszennyező, vagy inhumánus módon történt (pl. gyermekmunka, állatkínzás stb.). Különböző szervezetek és egyének nem győzik a médián és a moralizáló mémeken keresztül a fogyasztóra hárítani az ezzel kapcsolatos felelősséget, holott

[17] Szűts Zoltán, „Az internetes források használata az irodalomtudományban,” *Irodalomismeret* 23, 4. sz. (2012), 71–76, `http://www.irodalomismeret.hu/files/2012_4/szuts_zoltan.pdf`, 72.

[18] Itt szeretném megköszönni egyik névtelen bírálóm megjegyzését, miszerint ez egyre inkább a *mainstream* részévé válik Magyarországon is; úgy érzem, ez is megerősíti témaválasztásom jogosságát.

[19] Ejtsd [bogsu] magyaros átírással [bokszu]: „bosszú” – a *K-drámá*k, főleg a *szagük*ok egyik gyakori eleme.

[20] Jean Baudrillard, „A szimulákrum elsőbbsége,” ford. Gángó Gábor, in *Testes könyv I,* szerk. Kiss Attila, Kovács Sándor et al. (Szeged: Ictus–JATE, 1981 [1996]), 161–193, 162.

[21] Vadai István, „Balassi Bálint elvegyült énekei,” *Irodalomtörténeti Közlemények* 118, 3. sz. (2014), 393–402, 394, `http://itk.iti.mta.hu/megjelent/2014-3/vadai.pdf`.

[22] Gondoljuk csak olyan művekre, mint a *Jekyll és Hyde* (Robert Louis Stevenson, 1886), *A gazdag szegények* (Jókai Mór, 1890), *Az éhes város* (Molnár Ferenc, 1901), *Budapest* (Kóbor Tamás, 1901), *Nagyvárosi fények* (Charlie Chaplin, 1931) stb.

– hasonlattal élve – nyilvánvalóan nem az emberek alkoholszomja, de még csak nem is az alkoholtilalom tette azzá Al Caponét, aki lett. Al Capone nagy valószínűséggel akkor is maffiózó lett volna, ha nem illegális alkoholkereskedelemből gazdagszik meg. Amikor a *cyberbullying* kérdését úgy közelítjük meg, hogy az áldozatokra fókuszálunk, és megpróbáljuk számukra relativizálni a történteket („ne olvasd a kommenteket," „ne foglalkozz vele" stb.), a perceptuális orvoslás hibájába esünk. Az agyi érzékelést kikapcsoló fájdalomcsillapító nem gyógyítja a sebet, a gyógyuláshoz ennél több kell.

Nap mint nap szállítja nekünk a média a gyűlölettel kapcsolatos történeteket, ami önmagában is képes arra, hogy szorongást keltsen bennünk. A történetek hatására akkor is szenvedést élhetünk meg, ha a bennünket körülvevő személyes fizikai térben történetesen egy teljesen harmonikus világban élhetünk. Átérezhetjük mások szenvedését, mint a jólétben élő Buddha, aki a négy gondolatébresztő látvány során szembesül az öregség, a betegség és a halál létezésével.

Evely Waugh hősének esete[23] – aki mindaddig egy általa harmonikusnak hitt világban él, míg el nem bocsátják, és ekkor ébred rá, hogy rajta kívül már mindenki más tudta, mi vár rá, mégsem akadt senki, aki közölte volna vele ezt a tényt – jól példázza, hogy biztonságosnak és szépnek hitt világunkban ugyanannyi rosszindulat áramolhat az ember felé, mint amennyit most az interneten keresztül megkap, csak a világháló nélküli (IRL) kommunikációban az egyszerűen nem jut el hozzá. A filozófia és a pszichológia területén is történtek törekvések abba az irányba, hogy majdan kezünkbe vehessük „a gyűlölet általános elméletét".[24] Egyelőre azonban még túl keveset tudunk arról, hogyan működik pontosan a gyűlölet, így kevéssé tudjuk kezelni. Úgy tűnik, sem a jog, sem az újságokban megjelenő pszichológiai tanácsok nem elegendőek ahhoz, hogy a hálózatos világban közelivé váló gyűlölettől megóvjuk magunkat és gyermekeinket.[25]

2019 elején az egész világot bejárta a dél-koreai K-pop sztár, Seungri és társai botrányának a híre.[26] A Seungri-botrányt voltaképpen nem lehet egyetlen esetnek nevezni, ugyanis nem egyetlen bűncselekményre, hanem bűncselekmények többékevésbé összefüggő hálózatára derült fény.[27] A cselekmények egy részében, az ún. Jung Joon-young-botrányban,[28] illetve abban, hogy a cselekmények napvilágra ke-

[23] Evelyn Waugh, *A megboldogult,* ford. Ottlik Géza (Budapest: Európa, 1957).

[24] Lásd *hate studies.*

[25] Fábián Tamás, „Instagram-szavazás miatt lett öngyilkos egy tinédzser Malajziában," *Index,* 2019. máj. 16., `https://index.hu/techtud/2019/05/16/instagram-szavazas_miatt_lett_ongyilkos_egy_tinedzser_malajziaban/`.

[26] „The Great Seungri Scandal, the biggest scandal to ever hit Korean entertainment," *The Asian Theory*, 2019. márc. 21., `https://www.youtube.com/watch?v=ztCBLzwedKE`; Kang Buseong, Song Jung-a and Edward White, „K-pop scandals spark soul-searching in South Korea," *Financial Times*, 2019. ápr. 21., `https://www.ft.com/content/3cd6072c-5137-11e9-b401-8d9ef1626294`.

[27] Éppen a tanulmány írásakor jelent meg egy hosszabb hírösszefoglaló magyarul a botrányról: Presinszky Judit, „Kígyózik a K-pop botrány: a népszerű klubok éveken át bedrogozott nőket futtattak," *Index,* 2019. máj. 24., `https://index.hu/kulfold/2019/05/24/del-korea_prostitucio_kereskedelem_k-pop_botrany_bigbang_seungri_li_szunghjon_yg_entertainment/`.

[28] Rövid megjegyzés a koreai nevekről: a legtöbb név háromszótagos, a vezetéknév (családi név) elöl áll, mint a magyarban. Ez általában egyszótagos, pl. „Kim". A nevek latinbetűs transzliterációja angol alapú, azonban a koreai és angol nyelvű médiában különbözik. A magyaros átírástól, mivel az koreai és egyéb nemzetközi oldalakon visszakereshetetlen, eltekintek. A személy kapott nevének

 DOI 10.31400/dh-hun.2019.2.387

rülhettek, nagy szerepe volt az úgynevezett információs és kommunikációs technológiáknak (IKT).[29] A botrány egyes értelmezői pusztán a dél-koreai társadalom jellegzetességeiben keresik az okokat,[30] de az ügy valójában ennél bonyolultabb.

A Magyarországon is sikerrel vetített dél-koreai sorozatok óta (pl. *A palota ékköve*[31] vagy *A császárság kincse*[32]) már nem szorul részletes ismertetésre, hogy mi az a *K-drama.* Psy 2012-es *Gangnam Style*-ja, a BTS[33] és a Blackpink[34] világsikere óta nem kell magyarázni, mi az a *K-pop.* A *Filmvilág*ból is hallhattunk a *koreai hullámról*, azaz a *hallyu*ról, más néven *Korean wave*-ről,[35] s valószínűleg az is sokak előtt ismert, hogy a *K-drama* és a *K-pop industry* létrejöttében mekkora szerepe volt a koreai[36] államnak.[37] Voltaképpen a nemzet mint *brand*[38] jelenik meg a koreai kulturális iparban. A gigasztárok árbevétele és rajongói tábora mellett eltörpül a Nyugat vezető sztárjainak teljesítménye is.[39] A pénzbevételek jelentős része azonban nem a sztároké, hanem az őket foglalkoztató koreai iparágé.[40] Ez utóbbi a sztárokkal mint még hírnévre aspiráló fiatalokkal köt szerződést, és részletesen szabályoz olyan kérdéseket is, mint az illető

(given name) átírásakor a kétszótagos név átírása vagy egyben (pl. „Joonyoung," „Jonghyun"), vagy külön („Joon Young," „Jong Hyun") vagy kötőjellel történik, ilyenkor a második szótag általában kis kezdőbetűs („Joon-young," „Jong-hyun"), de létezik kötőjeles nagy kezdőbetűs forma is („Joon-Young," „Jong-Hyun"). Én a koreai angol nyelvű médiában megszokott formát használom a nevek átírásakor, ill. szerzőre való hivatkozáskor az adott írásban szereplő alakot.

[29] Koreai példát választottam a személyiségi jogok infokommunikációs eszközökön keresztül történő megsértésére, azonban köztudottan másutt is fordulnak elő hasonló esetek, gondoljunk például az úgynevezett *bosszúpornó* műfajára. A kukkoló kamerák *(molka)* illegális tartalmának fogyasztói az egész világon megtalálhatók. Ugyanígy, a tárgyalt bűncselekmény-sorozat említett elkövetői férfiak, ám pl. a *molká*k nyilvános női mellékhelyiségekben történő elhelyezését valószínűleg nők végezték.

[30] Yonden Lhatoo, „From Seungri to Jung Joon-young: South Korea's toxic masculinity explained," *South China Morning Post,* 2019. ápr. 7., `https://www.youtube.com/watch?v=u6OWB2-Q92s`.

[31] 대장금 (Dae Jang Geum / Jewel in the Palace) 2003, rend. Lee Byung-hoon.

[32] 기황후 (Empress Ki) 2013–2014, rend. Han Hee és Lee Sung-joon.

[33] 방탄소년단 (Bangtan Sonyeondan, 'golyóálló cserkészek') – 2013-ban alapított, héttagú dél-koreai fiúegyüttes.

[34] 블랙핑크 – 2016 óta működő, négytagú dél-koreai lányegyüttes.

[35] Teszár Dávid, „Az elfeledett háború," *Filmvilág*, 4. sz. (2008), 28–31, `http://filmvilag.hu/xereses_frame.php?cikk_id=9315`, Teszár Dávid, „Kistigrisből nagy tigris," *Filmvilág*, 5. sz. (2016), 30–34, `http://filmvilag.hu/xereses_frame.php?cikk_id=12729`.

[36] A koreai jelzőt dél-koreai értelemben használom itt és a továbbiakban.

[37] William Tuk, „The Korean wave was deliberately created by the Korean government: Korean entertainment companies used the popularity of Korean movies and dramas in other countries to successfully export k-pop" in William Tuk, *The Korean Wave: Who are behind the success of Korean popular culture?* (Leiden: Leiden University History of European Expansion and Globalization MA, Supervisors: Prof. dr. Jos Gommans, Mrs. drs. Monique Erkelens, 2012), 48.

[38] Lee Kyung-Mi, „Toward Nation Branding Systems: Evidence from Brand Korea Development," *Journal of International and Area Studies* 18, 1. sz. (2011), 1–18, `https://www.jstor.org/stable/43111488`.

[39] „The Great Seungri Scandal, the biggest scandal to ever hit Korean entertainment," *The Asian Theory*, 2019. márc. 21., `https://www.youtube.com/watch?v=ztCBLzwedKE`.

[40] „How much money a K-pop idol makes (according to a former K-pop idol)," *SBS PopAsia HQ*, 2018. jan. 29., `https://www.sbs.com.au/popasia/blog/2018/01/29/how-much-money-k-pop-idol-makes-according-former-k-pop-idol`.

sztár magánélete.[41] Világszerte, így Magyarországon is megjelentek azok a hírek, melyek alapján egyes sztárok öngyilkossága a sztárgyár embertelen bánásmódjával van összefüggésben.[42]

Dél-Koreában nemcsak a prostitúció, hanem a pornográfia is törvénybe ütköző.[43] Egyesek részben ennek tudják be, hogy olyan méreteket öltött az országban a titkos, erotikus vagy pornografikus célú, kukkoló rejtett kamerák *(molka)* telepítése, hogy a téma állandóan napirenden van.[44] A *molka*-ellenes tüntetések jelszava: „My life is not your porn."[45] A hatóságok szinte szélmalomharcot vívnak az elkövetőkkel, az elnök külön nyilatkozatban foglalkozott a kérdéssel,[46] aktivisták harcolnak az áldozatok jogaiért,[47] a külföldieknek szóló, angol nyelven sugárzó *Arirang* televízió társadalmi célú reklámokat sugároz a „szexuális zaklatástól mentes" Korea jelszavával,[48] illetve vannak olyanok, akik hivatásszerűen *molka*-mentesítik a szállodákat, moteleket, nyilvános illemhelyeket, stb.[49] Ezeknek a jogsértő módon készített intim felvételeknek a terjesztése nem korlátozódik az ország határain belülre, vagyis a készítő bűnelkövetők nem pusztán személyes aberrációik kielégítésére telepítik a molkákat, hanem ipari jelleggel szolgálnak ki különböző, erre szakosodott hálózati oldalakat, adatbázisokat.[50]

A különböző, befolyásos férfiakat illegális prostitúcióval kiszolgáló, Seungri nevével fémjelzett, gangnami *Burning Sun* lokálból induló botrány rövid idő alatt Seungri baráti körét is elérte, és egy másik K-pop sztár, Jung Joon-young körül tetőzött.

[41] James Griffiths, „Can K-Pop stars have personal lives? Their labels aren't so sure," *CNN Entertainment,* 2018. szept. 22., https://edition.cnn.com/2018/09/21/entertainment/kpop-dating-hyuna-edawn-music-celebrity-intl/index.html.

[42] Kadi-Maarja Võsu, „K-popi sünge varjukülg," *Postimees*, 2018. júl. 1., https://elu24.postimees.ee/4510844/k-popi-sunge-varjukulg; Föld S. Péter, „Öngyilkos lett az egyik legismertebb dél-koreai popsztár," *FüHÜ,* 2017. dec. 19., https://fuhu.hu/ongyilkos-lett-egy-del-koreai-popsztar/; „This Korean Actress Is Risking Her Life To Expose The Truth About Jang Ja Yeon," *Asian Boss,* 2019. ápr. 21., https://www.youtube.com/watch?v=8swU0C0U4ec.

[43] Choe Sang-Hun, „South Korean Court Upholds Ban on Prostitution," *The New York Times,* 2016. márc. 31., https://www.nytimes.com/2016/04/01/world/asia/south-korea-upholds-prostitution-ban.html; H. M. Kang, „South Korea Bans Access to Porn Sites," *The Korea Bizwire,* 2019. febr. 20., http://koreabizwire.com/south-korea-bans-access-to-porn-sites/132882.

[44] Victoria Kim, „K-pop's K-porn problem: Growing scandal highlights South Korea's spy-cam epidemic," *Los Angeles Times,* 2019. ápr. 3., https://www.latimes.com/world/asia/la-fg-kpop-porn-scandal-20190403-story.html.

[45] 'Az én életem nem a te pornód.' Emily Pacenti, „My Life Is Not Your Porn: The Korean Social Movement You Haven't Heard Of," *Spire Magazine,* 2019. márc. 12., https://spiremagazine.com/2019/03/12/my-life-is-not-your-porn-the-korean-social-movement-you-havent-heard-of/.

[46] Justin McCurry, „Spycams, sex abuse and scandal: #MeToo reaches Korean pop," *The Guardian,* 2019. márc. 22., https://www.theguardian.com/music/2019/mar/22/metoo-k-pop-music-industry-sexaul-assault-scandals-korean-cultural-life.

[47] „Trailblazers: Fighting South Korea's spy cam porn," *BBC News,* 2018. dec. 3., https://www.youtube.com/watch?v=bHNTnaOR_YU.

[48] Ez a reklám a belföldi csatornák műsorán koreai nyelven nem szerepel – legalábbis tapasztalataim szerint nem.

[49] „Korea's No.1 Spy-Cam Hunter On A Mission To Stop Illegal Porn," *Asian Boss,* 2019. ápr. 1., https://www.youtube.com/watch?v=xGpDp86RY5s.

[50] Uo., illetve Kim, Victoria: „K-pop's K-porn problem."

 DOI 10.31400/dh-hun.2019.2.387

Amit a médiából tudni lehet, az röviden annyi, hogy az 1989. februári születésű Jung Joon-young (JJY)[51] a telefonján 2015–2016-ban használt egyik *KakaoTalk*-os[52] chatszobából[53] elfelejtett kilépni, így az és annak tartalma nem törlődött automatikusan, hanem megmaradt akkor is, amikor a Seungri-botrány szálai JJY-hoz vezettek.[54] A chatszobában JJY általa illegálisan felvett pornografikus tartalmakat osztott meg, melyek azt ábrázolják, ahogy drogokkal cselekvésképtelenné tett nőkkel létesít szexuális kapcsolatot. JJY barátai, akik a chatszoba tagjai voltak, a videókat megnézték és azokhoz trágár, obszcén, a videókon szereplő nőket emberi méltóságukban sértő, vagy a nőket általában mint szexuális tárgyakat kezelő kommenteket fűztek, s nem mellesleg nem szólaltak fel és nem léptek fel JJY többszörösen jogsértő tevékenysége ellen.

A JJY chatszobájának tartalmát a nyilvánossággal megismertető újságírónő, Kang Kyung-yoon, az *SBS funE* riportere egy interjúban kiemeli, hogy egy olyan bűncselekmény-sorozatról van szó, melyben az elkövetők híresség, az áldozatok pedig hétköznapi emberek voltak.[55] A fenti kijelentés pontos értelmezéséhez nem árt tudni, hogy Koreában létezik a „szélsőséges rajongó"-ként magyarra fordítható úgynevezett *sasaeng*-jelenség; a *sasaeng*ek által zaklatott híresség ekről szóló hírek sajnálatosan gyakoriak.[56] Ez az eset tehát, mint mondja, azért különleges, mert itt nem a híresség et zaklatták, hanem a híresség ek voltak a zaklatók, illetve elkövetők.

Március 14-én két másik K-pop csillag, az 1989. decemberi születésű Yong Jun-hyung (YJH) és az 1990. márciusi születésű Choi Jong-hoon (CJH) ismerte el, hogy ők is megnézték és obszcén kommentekkel illették JJY videóit, egyúttal bejelentették visszavonulásukat a rivaldafényből. CJH-ról áprilisban kiderült, hogy ő maga is készített hasonló videókat nőkről, illetve számos más bűncselekményben érintett.[57] Egy nappal később egy újabb sztár, a kötelező sorkatonai szolgálatát töltő, 1990. májusi születésű

[51] Mivel a magyar olvasók számára nehézséget okozhat az egymáshoz hasonló koreai nevek elkülönítése, néhol csak monogramokkal fogok utalni az egyes személyekre.

[52] Koreában népszerű csevegőalkalmazás.

[53] A Koreában használt alkalmazásokról alapos áttekintést nyújt Yoo Jinil, „A koreai információs társadalom kultúra-specifikus jelenségei," *Infonia. Információs Társadalom* 14, 3. sz. (2014), 28–43, `https://epa.oszk.hu/01900/01963/00045/pdf/EPA01963_informacios_tarsadalom_2014_3_28–43.pdf`. Hasznos olvasmány lehet továbbá a digitális bölcsészet iránt érdeklődőknek Szűts Zoltán, „Információs társadalom Koreában – úton a teljes behálózottság felé," *Infonia. Információs Társadalom* 14, 3. sz. (2014), 5–18, `https://epa.oszk.hu/01900/01963/00045/pdf/EPA01963_informacios_tarsadalom_2014_3_05–18.pdf`.

[54] R. Jun, „Lawyer Explains How Jung Joon Young Unwittingly Saved Chatroom Conversations For Police," *Soompi,* 2019. ápr. 12., `https://www.soompi.com/article/1316777wpp/lawyer--explains-how-jung-joon-young-unwittingly-saved-chatroom-conversations--for-police`.

[55] yckim124, „Reporter Kang Kyung Yoon talks about her anger and shock after discovering Jung Joon Young's group chat," *Allkpop,* 2019. ápr. 15., `https://www.allkpop.com/article/2019/04/reporter-kang-kyung-yoon-talks-about-her-anger-and-shock-after-discovering-jung-joon-youngs-group-chat`.

[56] Lásd pl. „13 Extremely Disturbing Stories Of Sasaengs That Went Too Far," *Koreaboo,* 2018. okt. 11., `https://www.koreaboo.com/lists/13-disturbing-stories-sasaeng-fans-went--far-1/`.

[57] Jelen szöveg szerkesztésének idején adta hírül a sajtó, hogy Jung Joon-young hat-, Choi Jong-hoon pedig ötéves börtönbüntetést kapott. Park Si-soo, „K-pop singer Jung Joon-young gets six years' jail

Lee Jong-hyun [LJH] ügynöksége nyilatkozott arról, hogy a sztár is megnézett, és a nőket sértő módon kommentelt videót a chatszobában.[58] Az 1993. júliusi születésű Roy Kimet és az 1990. novemberi születésű Eddy Kimet illegális videofelvételek terjesztésével vádolják, mivel továbbosztottak illegális, a netről letöltött videókat. A színész Cha Tae-hyun (szül. 1976. márc.) és a komikus Kim Jun-ho (szül. 1975. dec.) illegális szerencsejáték-tevékenységére is a már említett csevegőprogramon keresztül derült fény – azóta mindketten visszavonultak.

Több más zenész, illetve modell mellett a rendőrség közlése szerint a visszavonulását bejelentő YJH és a sorkatonai szolgálatát töltő LJH egyelőre csak tanúként szerepel az ügyben.[59] Lee Jong-hyun (LJH), illetve együttese, a CNBLUE rajongóit is sokkolta a hír, s a rajongótábor gyakorlatilag kettészakadt. A rajongók egyik része azért kezdett internetes aláírásgyűjtésbe, hogy a bálvány, akiben csalódtak, távozzon a csapatból, a rajongók másik része pedig azért indított online petíciót, hogy a sztár, aki még az eset történtekor kilépett a chatszobából, maradjon a zenekar tagja, és tekintsék az egészet ifjúkori ballépésnek.

Hónapok óta zajlik a kommentháború, a gyűlölködők miatt Lee Jong-hyun instagramos profiljáról még március 15-én (ő vagy a menedzsmentje) minden tartalmat töröltek, de a zenekar oldalán, illetve a botránnyal kapcsolatos sajtóhírekben, valamint a sztárhoz és együtteséhez kapcsolható különböző rajongói oldalak, közösségimédia-profilok alatt is folyamatosak a szélsőségektől sem mentes megnyilvánulások. Lee Jong-hyun, aki várhatóan 2020. március 25-én szerel majd le a katonaságtól, bár reménykedhet abban, hogy addigra talán enyhül valamelyest a rajongók haragja, valószínűleg nem bizakodhat a teljes megbocsátásban.[60] Hogy miért nem, annak számos oka van.

A kommentekben rajongói közt a legmegbocsátóbbak a dél-amerikaiak és a japánok voltak, a legengesztelhetetlenebbnek pedig a dél-ázsiai muszlim országokból való rajongók tűnnek. (Ez természetesen nem jelenti azt, hogy minden, az adott régióból származó rajongó ugyanazt gondolja, ez csupán egy jól kivehető tendencia a különböző kommentfolyamokban.)

A CNBLUE a többi K-pop együttessel ellentétben nem Koreában, hanem Japánban debütált, azóta is számos lemezt adtak ki japánul, és rendszeresen visszatérnek koncertezni az országba. Bár az ő esetükben is nagy szerepe van a tagok külső, fizikai megjelenésének, mint a klasszikus K-pop együttesek esetében, s az együttes mind a

for gang rape," *The Korea Times*, 2019. nov. 29., http://www.koreatimes.co.kr/www/art/2019/11/398_279522.html.

[58] Érdekes véletlen, hogy éppen annak a sztárnak az ügynöksége nyilatkozott március 15-én, aki korábban *A maszkos énekes* című tévéshow-ban egy bűvös kockát formázó maszkban versenyzett.

[59] „Roy Kim suspected of spreading illegally taken photos, booked by police," *SBS PopAsia HQ*, 2019. ápr. 4., https://www.sbs.com.au/popasia/blog/2019/04/04/roy-kim-suspected-spreading-illegally-taken-photos-booked-police.

[60] 2019. augusztus 28-án jelentette a koreai média Lee Jong-hyun távozását ügynökségétől, az FNC Entertainmenttől, s egyúttal kilépését a CNBLUE együttesből. J. K., „Breaking: Lee Jong Hyun Announces Departure From CNBLUE," *Soompi*, 2019. aug. 28., https://www.soompi.com/article/1348668wpp/lee-jong-hyun-announces-departure-from-cnblue. Mivel az együttes neve az ügynökség koncepciója szerint a C(ode) N(ame): B(urning), L(ovely), U(ntouchable), E(motional) szavak alkotta mozaikszó, ahol Jong-hyun adta a B-t, a „Burning"-et, hosszabb távon feltételezhető, hogy egy új énekes-gitáros-zeneszerzővel helyettesítik majd a csapatban.

 DOI 10.31400/dh-hun.2019.2.387

négy tagja színészként is szerepelt már különböző tévésorozatokban, ahogy a K-pop sztárjainál általában megszokhattuk, a CNBLUE nem igazán tipikus K-pop-csapat. Az ő zenéjük inkább rock vagy pop-rock, mint klasszikus értelemben vett K-pop. Az együttes tagjai nem énekes-táncosok, hanem zenészek és énekesek, akik számaik nagy részét maguk írják, és a zenét is maguk szerzik. A kommentelők egy része szerint, mivel rajongótáboruk jelentős része Japánban van, ahol – mint az köztudott – a pornográfia nem tilos, (mint Dél-Koreában vagy a dél-ázsiai muszlim országokban) kevésbé tekintik bűnnek a rajongók azt, ami Lee Jong-hyun (LJH) chatszobabeli tevékenységéről kiderült. A magam részéről nem hiszem, hogy cselekedetével azért megengedőbbek valahol, mert ott nem tilos a pornográfia, hiszen ott, ahol a pornográfia legális, csakis konszenzusos módon az, még ha nehéz is kideríteni bizonyos esetekben, hogy konszenzusról vagy kényszerítésről volt-e szó. A videót készítő és megosztó JJY tevékenysége mindenütt bűncselekmény lenne.

A K-pop sztárjai között, akik mindig tökéletes sminkben és ruhában jelentek meg a médiában, Lee Jong-hyun üdítő kivételnek számított, mivel mert hétköznapi módon nemtörődöm lenni. Eredeti, néha talán eredetieskedő figurának tűnt. És általában a chatszoba-botrány minden érintettjéről el lehet mondani, hogy európai szemmel teljesen érthetetlen, hogy jóképű és népszerű fiatal férfiak miért ilyen módon élik ki szexuális vágyaikat. A kommentelők egy része szerint ebben a koreai popzenei iparnak az a gyakori szerződési kitétele bűnös, mely szerint a sztárok nem randevúzhatnak és nem élhetnek párkapcsolatban a szerződés ideje alatt, és a rájuk kényszerített cölibátus miatt jelentkezik perverziókban az amúgy teljesen természetes szexuális vágy. A kommentelők más része szerint a cölibátusnak ehhez semmi köze, mert ezek az emberek romlottak, az egész csak a hatalomról szól, és ugyanígy viselkednének akkor is, ha szabadon randevúzhatnának és köthetnének párkapcsolatot.

Nem lehet nem összevetni a történteket a Nyugaton nagyjából ugyanekkor hasonló botrányt kavaró Michael Jackson-filmmel, amely újra elővette a néhai sztár állítólagos pedofíliájának vádját. S a K-pop világában is felmerült a kérdés, vajon rajonghatunk-e valakinek a műveiért, zenéjéért, ha az illetőről kiderül, hogy magánemberként bűncselekményeket követett el. Nem hagyhatjuk figyelmen kívül azt sem, hogy LJH-t is, amíg meg nem vádolják valamivel és a bíróság el nem ítéli, megilleti az ártatlanság vélelme. A követőinek és a kommentelőknek csak feltételezései lehetnek arról, hogy „feltehetően", „a nyilvánosságra hozott adatok szerint" mit érezhetett és gondolhatott, amikor a chatszobában rögzített dolgokat írta és tette. Szenvedést mérünk össze szenvedéssel: a megerőszakolt, megalázott nőkét, a perverz sztárét, amikor a dolog kiderült, és a csalódott rajongókét, akik megundorodtak bálványuktól vagy kedvencüktől. A hálózat adta lehetőségnél fogva közvetlenül, azonnal és van, hogy keresetlen formában is hangot adhat a hírfogyasztó a véleményének.

Az igazságszolgáltatás különböző szerveit azért tartjuk fenn, hogy helyettünk szerezzenek szakértelmet a tárgyban, és ennek segítségével intézkedjenek helyettünk és az érdekünkben. A hálózat, vagyis a felhasználók tömege viszont képes *virtuális pellengér*re állítani gyakorlatilag bárkit. Vajon eljön valamikor az az idő a világháló korában, amikor a felhasználók újra közvetlen módon gyakorolják a demokráciát, mint az ókori Athénban, és nem lesz szükség többé képviselőkre?

LJH és YJH bűne, ami gyűlölködő kommentek formájában fejükre száll: bűnös úton előállított tartalom *fogyasztás*a, és dehonesztáló *kommentelés*e. Ők a tartalomfogyasztók, akik fogyasztásuk révén bűnrészessé válnak valamilyen szinten az általuk fogyasztott tartalom miatt a tartalom előállításában. Természetesen nem a jogszabályok, hanem más tartalomfogyasztók szerint; ahogy azt látjuk, hogy egyes környezetvédelmi kampányok vagy személyek megbélyegzik azokat, akik nejlonszatyrot vagy más egyszer használatos műanyagárut vesznek igénybe, holott nyilvánvalóan nem az egyén felelős az általa használt nejlonszatyornak sem az előállításáért, sem a szeméttelep utáni sorsáért. LJH és YJH az átlagfogyasztó megtestesítői. A fogyasztó szórakozásvágya és kíváncsisága tartja fenn a kattintásvadász oldalakat, a bulvármédiát és persze a pornográfiát is. A pornográfia (mint műfaj) viszont jogilag nem azonos magatehetetlen nők megerőszakolásáról készített illegális felvételekkel, ahogy a bulvármédia sem paparazzifotók publikációs fóruma. Ezeket a videókat nem LJH és YJH rendelték meg, a videók létrejöttéért nem mondhatók jogilag felelősnek. A kommentelők egy része szerint ezekre a privát körben kapott videókra az lett volna a megfelelő reakció, ha feljelentik a barátaikat a rendőrségen, mások szerint viszont reakciójuk esendő emberi reakció volt.

A népi bölcsesség kommentek formájában történő megnyilvánulása persze lehet nagyon szórakoztató; számos ember olvas szórakozásból kommenteket. A magyar online médiában szinte mémmé vált „sunáznám" is egészen addig nagyon mókás lehet, amíg nem rólunk szól, nem ránk vonatkozik. Az online médiában való jelenlétünkön keresztül (és itt most a közösségi médiát a fogalom részének tekintem) állandóan megítélünk és megítéltetünk, sőt, maga a média éppen erre buzdít, hogy a lájkvadászaton és a diszlájkkerülésen, s a követőink számán keresztül állandóan megméressünk és pénzben kifejezhető értékké váljunk. Nem sokkal vagyunk többek ezzel, mint a *Mátrix*[61] elememberei. Nem a *tartalom* számít, hanem a *forgalom*.

Belépek a közösségi oldalamra, nézem a hírfolyamot. Idősebb asszonyok egy fotón népviseletben. *Like.* Semmit nem tudok róluk. Jó emberek vagy rosszak? A fényképük képvisel valamit, amivel *azonosul*nom illik. De a lájkomat úgy olvassák-e, hogy azzal azonosultam, amit én gondoltam, vagy valami teljesen mást látnak bele? A rengeteg lájk, amit kiosztottam, vajon kinek-minek szólt? A másodperc törtrésze alatt kiosztjuk a lájkjainkat. Minden eddiginél felszínesebb ítéleteket kell hoznunk. Reakcióink visszakereshetőek, lájkjaink, kommentjeink alapján megítélhetőek vagyunk, mint LJH, amikor azt írta, hogy „küldhetne a tesó neki is egy p…-t, amit az még nem b…-tt meg."[62]

Nagyon jól mutatja a külső és belső értékrend kétarcúságát, hogy a privát chatszobában megosztott videókkal JJY nyilvánvalóan „menőzni akart a haverjai előtt"; cselekedeteinek napvilágra kerülésekor viszont nem egyszerűen a feléje irányuló gyűlölethullámmal kellett szembesülnie, és átélnie a megszégyenülést, hanem valóban éreznie kellett a *szégyen*t viselkedése miatt. A távol-keletiek közismert udvariasságát a Nyugat régebben alázatosságnak hitte, holott erről szó sincs. Az udvariasság olyan

[61] *The Matrix [Mátrix]* (1999), rend. The Wachowskis.

[62] Hanan Haddad, „CNBLUE's Lee Jong-Hyun Reportedly Entangled In »Secret Porn« Scandal," *Eonline,* 2019. márc. 15., `https://www.eonline.com/ap/news/1023879/cnblue-s-lee-jong-hyun-reportedly-entangled-in-secret-porn-scandal`.

 DOI 10.31400/dh-hun.2019.2.387

viselkedési forma, mely lehetővé teszi mindkét fél számára, hogy *elkerülje a szégyent.*[63] A távol-keleti viselkedéskultúra szerint a szégyen az egyik legnagyobb rossz, ami az emberrel történhet; és természetesen a másik felet megszégyenítő viselkedés is szégyen. A koreai tévédrámákban ezért gyakori, hatásos gesztusok például a másik ember megbilincselt kezének eltakarása mint a szégyentől való megóvás gesztusa, vagy a térden állva való könyörgés motívuma mint a szégyen önkéntes vállalása a cél érdekében.

Bűncselekmények esetében helyes, ha a rendőrség minden adathoz hozzáfér, ami segít tisztázni az ügyet. Biztos azonban, hogy nem helyes a jog előtt bűnelkövetőnek nem minősülő személyek magánjellegű megnyilvánulásait nyilvánosságra hozni. Akkor sem helyes (vagy talán kifejezetten akkor nem helyes), ha az illető sötét oldalát leplezzük le. Vagy eljön az az idő, amit a kevéssé értékelt *A kör*[64] című film felvázol, ahol minden lépésünket önként osztjuk meg mindenkivel, hogy ez az áttetszőség és ellenőrizhetőség védjen meg bennünket? Orwell disztópiájában a totális megfigyelés egy diktatúrában valósul meg, *A kör* viszont egy olyan világot rajzol elénk, ahol a közösség tagjai végül nem vezetői nyomásra, hanem önként válnak „átlátszóvá," s ezen keresztül biztosítják, igazolják testvériségüket, és védik magukat. Az mindenesetre ma már elvárásként jelenik meg, hogy folyton elérhetőek legyünk – mint Pléh Csaba írja –, „az állandó készenlét világában kell élnünk."[65] Ha nem vagyunk folyamatosan elérhetőek, az sértésnek számít, vagy arra utalhat, hogy éppen tilosban járunk – legalábbis a közvélekedés szerint.[66]

Közszereplők esetében világszerte nincs konszenzus azt illetően, hogy mely adatok tartoznak a nyilvánosságra, és melyek nem. A papírkönyves világban kedves adat volt, hogy mikor van a sztár születésnapja. Ma annyi személyes adat érhető el a neten mindannyiunkról, hogy azokkal nagyon könnyű visszaélni.[67]

A Távol-Keleten a különböző sztáradatbázisokban nemcsak az illető születésére és munkásságára vonatkozó adatokat találjuk meg, hanem a vércsoportját is.[68] A vércsoport ott (ezek szerint) nem minősül szenzitív adatnak; s mivel a különböző vércsoportokhoz ugyanúgy kötnek különféle személyiségjellemzőket, mint a csillagjegyekhez,

[63] Boyé Lafayette De Mente, *The Korean Mind: Understanding Contemporary Korean Culture* (Tokyo–Rutland, Vermont–Singapore: Tuttle Publishing, 2017); magyar fordításban: Boyé Lafayette De Mente, *A koreai észjárás: Ismerkedés a kortárs koreai kultúrával,* ford. Rohonyi András (Budapest: Pallas Athéné Könyvkiadó, 2018). A könyv részletes bevezetőt ad a koreai gondolkodás alapvető fogalmaiba. Jelen téma szempontjából lásd az alábbiakat: *chae-myeon* (a magyar fordításban: „mindenkinek meg kell őriznie az arcát," [2017]: 29–31; [2018]: 50–52); *changpi* („a szégyen kultúrája," [2017]: 33–34; [2018]: 58–59); *mangshin* („a szégyen elkerülése," [2017]: 241–242; [2018]: 356–357.)

[64] *The Circle [A kör]* (2017), rend. James Ponsoldt.

[65] Pléh Csaba, „A webvilág kognitív következményei, avagy fényesít vagy butít-e az internet," *Korunk* 22, 8. sz. (2011), 9–19, 13. `http://epa.oszk.hu/00400/00458/00571/pdf/EPA00458_korunk_2011_08_009–019.pdf`.

[66] Ez a folytonos készenlét, az állandó rendelkezésre állás természetesen akadálya az elmélyült munkavégzésnek is. Lásd Pléh, „A webvilág," 13.

[67] Mint a '90-es évek egyik filmsikerében (magyarul *A hálózat csapdájában* címet kapta), ahol Angela Bennett karakterlopás áldozata lesz. Egy bűnözői hálózat megpróbálja őt eltenni láb alól, s ennek részeként egyvalaki ellopja a személyazonosságát. *The Net [A hálózat csapdájában]* (1995), rend. Irwin Winkler.

[68] Lásd pl. `http://asianwiki.com`; `https://channel–korea.com`.

a rajongók álmodozhatnak a vércsoport ismeretében arról, hogy mennyire illenének össze a sztárral.

A rajongók itt (európai szemmel) egészen meglepő tárgyakat is vehetnek a sztár portréjával díszítve. Nemcsak sztáros poszterekkel és naptárakkal dekorálhatják ki a szobájukat, hanem vehetnek például életnagyságú kispárnát is a sztár képével. A fénykép természetesen: birtokbavétel, ahogy Susan Sontag írja.[69] A sztár fotójával ellátott kispárna a birtokbavétel szimbolikus aktusa, hiszen fizikai kontaktust tesz lehetővé az egyébként elérhetetlen személy képmásával. Nyilvánvalóan kellő gátlástalanság, vagy ha úgy tetszik, megszállottság szükségeltetik ahhoz, hogy valaki egy ilyen párnával bújjon ágyba. A sztáros kispárna már csak egy lépésre van a sztárra hasonlító szexbábutól. Ezek a párnák nyilvánvaló bizonyítékai a *sexploitation*nek, a filmsztárok tárgyiasításának és szexuális vonzerejük pénzre váltásának. A párnát megvásároló rajongó nyilvánvalóan nem tekinti önálló akarattal rendelkező individuumnak rajongása tárgyát, ilyen értelemben bizonyos mértékig mondhatni szexuális erőszakot követ el rajta. A különböző rajongói csoportok pedig egyenesen úgy definiálják magukat, mint az adott sztár „feleség"-ei.

A koreai sztárságban ugyanis a férfiaké a főszerep. A *hallyu*-sztárok iránti extatikus rajongás ahhoz hasonlítható, ami Nyugaton az '50-es, '60-as években történt, amikor a női rajongók különböző férfisztárok (mint pl. Elvis vagy a Beatles-tagok) látványától félőrülten sikítoztak és szó szerint ájuldoztak. Több *hallyu*-sztárnak van sokmilliós női rajongótábora a környező országokban, és számos közönségtalálkozós videó található a videómegosztókon, ahol jól láthatóan és hallhatóan a női közönség dominál. Maguk a közönségtalálkozók pedig olyan bizarr részleteket is tartalmaznak, mint például a sztár ruhát cserél egy átvilágított paraván mögött, míg a rajongók sikítozva nézik bálványuk meztelen testének árnyképét.

Nem alap nélkül éri az a kritika a koreai szórakoztatóipart, hogy prostituálják a sztárjaikat a közönség felé. Éppen a LJH-t érő gyűlölködő kommentekkel kapcsolatban jegyezte meg az egyik kommentelő, hogy álságos dolognak tartja, hogy éppen azok a rajongók ítélik el a sztárt egyetlen szexista kommentje miatt, akik hasonlóan szexista, és a sztárt szexuális tárgynak tekintő kommenteket fűznek nap mint nap annak képeihez és videóihoz.

Szerelem, 사랑해[70]

„Szimulálni annyi, mint úgy tenni, mintha lenne az, amink nincs."
(Jean Baudrillard: *A szimulákrum elsőbbsége*)[71]

A rajongók nyilvánvalóan abba a képbe képzelik szerelmesnek magukat, amit a sztárról a szerepei, megnyilatkozásai, valamint a hírek és a pletykák alapján kialakítanak.

[69] Susan Sontag, „A képvilág," in Susan Sontag, *A fényképezésről,* ford. Nemes Anna (Budapest: Európa Kiadó, 1981), 176–179, `http://mek.niif.hu/00100/00125/00125.pdf`.

[70] Ejtsd [saranghae] – magyaros átírással: *szaranghe* „szeretlek": ezt kiabálják a sztároknak a rajongók (külföldiek is).

[71] Baudrillard, „A szimulákrum elsőbbsége," 162.

 DOI 10.31400/dh-hun.2019.2.387

Klasszikus félreértés, hogy összetévesztik a sztár által megformált karaktert a sztárral. A *Faith* (신의, más angol címmel: *The Great Doctor*)[72] című *sageuk* (사극 – koreai, történelmi tévésorozat) rajongói nem Lee Min-hóba, a koreai szupersztárba lesznek szerelmesek, és nem is a történelmi Choi Young (최영, 1316–1388) tábornokba, hanem a Lee Min-ho által a *Faith*-ben megformált Choi Youngba. Egy többszörösen nem létező figura iránt alakítanak ki romantikus érzelmeket, de azt hiszik, Lee Min-hót szeretik. Nem kell ahhoz egy felvételnek pornografikusnak lennie, hogy intim közelségbe kerüljünk a képmással. A képmás, a szimulakrum egyrészt mindig tökéletes, másrészt amit mond, ahogy néz és ahogy viselkedik, mind olyan ideált testesít meg, mely hatással van a nézőre, aki nemcsak azt felejti el, hogy amit lát, fikció, hanem azt is, hogy illúzió, aminek a megszületése több száz ember munkája. A néző nem azért felejti el ezt, mert feltétlenül ostoba, hanem azért, mert a képek, a hangok és a mozdulatok nem az észre, hanem az érzékekre hatnak. Ráadásul a szimulakrum tetszés szerint kibe kapcsolható vagy végteleníthető.[73]

Tudjuk, hogy nem csak a filmek és a képek bolondítják meg az embereket. Klasszikus példákat ismerünk a fikció tudatmódosító hatására az irodalomból. A romantikus szerelem népszerű kritikusa, Alain de Botton[74] a könyvek hatására szerelemfüggővé váló Bovaryné gondolatait azzal szemlélteti, hogy az kvázi a mérleg egyik serpenyőjébe teszi a férje által biztosított romantikamennyiséget, a másikba pedig a szerelmesregényekből megismert romantika mennyiségét, s bizony joggal érzi úgy, hogy a mérleg nincs egyensúlyban. Bovaryné viszont nem arra a felismerésre jut, mint Alain de Botton, hogy a szerelmesregény mese, hazugság, fikció, ezért aztán Emma Bovary a tettek mezejére lép, amellyel mindannyiuk életét tönkreteszi.

Annak ellenére, hogy a *Bovaryné* a szerelmesregények szatirikus kritikája, az értelmezési hagyományban is létezik mind e regénynek, mind pedig a 19. század nagy *adultera*-regényeinek egy Alain de Botton által kritizált értelemben „romantikus"-nak nevezett olvasata. E szerint a nők a romantikára mint jogos jussukra formáltak igényt, de az Adyval szólva „durva kezek" nem adták ezt meg nekik. Még Móricznál is ez a logika mozgatja *Az Isten háta mögött* cselekményét. Van viszont a magyar irodalomban egy nagyon érdekes regény a romantika korából, ahol egy fiatalasszony Bovarynéhoz hasonlóan az olvasmányélményei hatására szeretné átélni a nagy romantikus szerelmet, melyre tökéletesen alkalmasnak is látszana egy, a környéken megjelenő titokzatos fiatalember. Ezt a regényt viszont nem lehet „romantikusan" olvasni, mert tele van *metafikciós, metanarrációs* humorral. Jósikának *Az első lépés veszélyei* című regényéről van szó. Véleményem szerint ez az igazi magyar *Bovaryné*, valódi magyar szatirikus ízzel. [75]

[72] *Shinui* (2012), rend. Kim Jong-hak.

[73] Lásd az *A. I.* című amerikai film végén az űrlények által a robotkisfiúnak adott szimulácót.

[74] Alain de Botton, *On Love,* 2016. júl. 10., `https://www.youtube.com/watch?v=v-iUHlVazKk`.

[75] Tóth Tünde, „A magyar Bovaryné-regény (humor, tragikum, narráció)," in *A regény és a trópusok: Második Veszprémi Regénykollokvium,* Veszprém: Veszprémi Egyetem Tanárképző Kar, Magyar Irodalomtudományi Tanszék, 2005. szeptember 29. – október 1.

Természetesen nem a 19. században kezdi el a fikció befolyásolni az olvasók viselkedését. A művészet mindig is hatott a befogadók viselkedésére.[76] Piccolomini 15. századi *De duobus amantibus*ának[77] hősnője is egy művelt fiatalasszony, aki olvasmányélményei hatására felismeri, hogy férje (azokhoz képest) milyen unalmas, ezért egy fiatal lovaggal kezd házasságtörő viszonyba. A latin nyelvű novella 16. századi, verses magyar fordítását röviden csak *Eurialus és Lucretia* néven emlegeti a szakirodalom, s ez az egyik legtöbbet elemzett széphistóriánk.[78] A széphistóriának a JJY-botrányt idéző jelenete, amikor Eurialus eldicsekszik a hódításával barátjának, s azt mondja, hogy szívesen megmutatná neki Lucretia meztelen testét. Azontúl, hogy a technológia mára sajnálatos módon megvalósította Eurialus vágyát, nyilvánvalóan mind Eurialus szavai, mind JJY szégyenletes tettei ugyanarról a tőről fakadnak.

Amióta pedig mozgókép létezik, létezik az őrületnek az a formája, hogy a sztárokkal tévesen azonosított szimulakrumokat jobban szeretjük, mint a hús-vér embereket körülöttünk. A szimulakrum iránt érzett szerelem természetesen azóta része az emberi kultúrának, mióta képesek vagyunk ezek előállítására. Erről szól a jól ismert mítosz is: Pygmalion, a szobrász beleszeret az általa készített szoborba.[79] A filozófia rávilágít arra, hogy a körülöttünk levő embereket is csak a tudatunk szűrőjén át ismerjük,[80] és csak az általunk róluk mentálisan megalkotott imágót tudjuk szeretni vagy utálni.

Ha az imágó változik, változik a hozzá való viszonyunk is: mint Lizzynek megváltozik a véleménye Mr. Darcyról a *Büszkeség és balítélet*ben, amikor annak kastélyába látogat. Lizzy úgy tesz, mintha nem a kastély miatt változott volna meg a véleménye, még viccelődik is ezen, de mégis ez a fordulópont kettejük kapcsolatában. Pontosabban az a pillanat, amikor Lizzy éppen a Darcyról készült festményt szemléli és közben a házvezetőnő kommentárjait hallgatja a férfiról. Lizzy nem Darcyba szeret bele, hanem a szimulakrumába, és nem Darcyt utálta, hanem azt, amilyennek ő képzelte a rendelkezésére álló információk (Darcy kezdeti elutasító magatartása és Wickham hazugságai) alapján. Maga a kastély tehát úgy is tekinthető, mint Darcy szimulakrumának szimbóluma.

Liza, a rókatündér története *amélie*-s bájjal[81] egyesítette az *Édes Anna*[82] és a *Shinigami no seido*[83] [kb. „a halálisten pontossága"] karaktereit egy fiktív „csudapesti" '70-es években. A film *férfilistája*[84] nyilvánvalóan rokonítja azt az ezredforduló férjvadászfilmjeivel, de Lizára mégsem „pasivadász szingli"-ként tekintünk, hanem mint egy

[76] Bovaryné mint női „Don Quijote" figurájához lásd: Fried István, „»Költőkirály« a(z anti-)modernitás antinómiái között," *Forrás* 31, 7–8. sz. (1999), 101–109, https://library.hungaricana.hu/hu/view/Forras_1999/?pg=746&layout=s.

[77] Aeneae Sylvii Piccolominei, *De duobus amantibus historia,* recensuit Iosephus I. Dévay, editio facsimile (Budapestini: Bibliopolis, 2008), http://www.bibliopolisz.hu/editiones/lucretia/devay/.

[78] B. Kis Attila és Szilasi László, „Még egyszer a Pataki Névtelenről: Történeti poétika és dekonstrukció, névtelenség és dialogicitás," *Irodalomtörténeti Közlemények* 96, 5–6. sz. (1992), 646–675.

[79] Kerényi Károly, *Görög mitológia*, ford. Kerényi Grácia (Budapest: Gondolat, 1977), 53–54.

[80] Lásd Platón barlanghasonlatát: *Az állam,* 516a-e. Magyar kiadása: Platón, *Az állam*, ford. Jánosy István (Budapest: Gondolat, 1988).

[81] *Le Fabuleux Destin d'Amélie Poulain [Amélie csodálatos élete]* (2001), rend. Jean-Pierre Jeunet.

[82] *Édes Anna* (1958), rend. Fábri Zoltán.

[83] *Suwîto rein: Shinigami no seido* (2008), rend. Kakehi Masaya.

[84] Amikor Liza lánc-randevúzik különböző férfiakkal.

 DOI 10.31400/dh-hun.2019.2.387

szimpatikus, bájos, naiv fiatal lányra. Nagyon érdekes azonban, hogy Liza a történet során egyáltalán nem szerelmes senkibe. Gyakorlatilag bárkit képes elfogadni, akiről azt hiszi, hogy szereti őt. Leszámítva Tomy Tanit, aki hiába epekedik érte évek óta. Amikor Liza egy könyvet tesz meg élete receptjének, forgatókönyvének, aminek a leírásait teljesítenie kell a boldogsághoz, tévedést követ el. A film fiktív világán belül Liza nem azért téved, mert egyébként irreális az elgondolás, hogy egy könyvben leírt eseménysor megvalósulását (vagyis a szimuláció realizációját) akarjuk megélni az életünkben. Liza tévedése a film belső logikája szerint abban áll, hogy az ő életéről nem az általa ronggyá olvasott romantikus ponyvaregény szól, hanem egy másik könyv: a szétvagdosott Rókatündér-katalógus.

Azt mondják, Zuckerbergék azért hozták létre a Facebookot, hogy válogatni tudjanak a lányok közt; maga a név is a modellügynökségek katalógusaira utal. Közösségi médiaprofiljaink több-kevesebb lelkesedéssel és szakértelemmel önmagunk (vagy menedzsereink) által legyártott szimulakrumaink, ahol olyannak mutathatjuk magunkat, amilyennek magunkat láttatni szeretnénk.[85] Ebben a közegben nemcsak a retusált instacelebek képein látható mozdulatok, gesztusok és szűrők merítik ki a pózolás fogalmát, hanem a nem-pózolás is pózként értelmeződik. A közeg felülírja, idézőjelbe teszi, relativizálja szándékainkat. „Mi mind egyéniségek vagyunk" – mantrázza a tömeg a *Brian életé*ben.[86]

A romantikus filmek hamis világábrázolása ellen sokan felszólaltak már; a széles közönséget megcélzó online önismereti tanácsadók is, mint például Attis (YouTube: *Tanulom magam*)[87] vagy a már említett Alain de Botton (YouTube: *The School of Life*).[88] Azontúl, hogy a romantikus filmek hazudta végzet (운명 – *unmyeong*)[89] az emberiség nagy többségének tapasztalatai alapján nem létezik, a technológiai fejlődés idővel nyilván lehetővé teszi, hogy nárcisztikus társadalmunk[90] megteremtse magának a tökéletes társat, vagy éppen gyereket a robotok személyében, mint az *A. I.* című filmben.[91] Mi lehet tökéletesebb szimulakrum, mint egy robot? A klasszikus *Metropolis*[92] óta a robotok nem ijesztő *gólem*ek, hanem érző lények, mint a romantikus irodalomban Frankenstein teremtménye, szerves alapú műembere Mary Shelleynél.

A koreai tévésorozatokban a szerelem kunderai lassúsággal[93] fejlődik ki. Erre kiváló teret ad a sorozati forma, hiszen nem kell mindent másfél-két órába sűríteni, mint

[85] Laila Koubia, „Yeey I'm on Facebook!:s," *Masters of Media,* 2010. szept. 27., `https://mastersofmedia.hum.uva.nl/blog/2010/09/27/yeey-im-on-facebook-s/`.

[86] *Life of Brian [Brian élete]* (1979), rend. Terry Jones.

[87] Attis, „Létezik a nagy Ő?" *TanulomMagam,* 2019. máj. 7., `https://www.youtube.com/watch?v=-BVu__bNONSk`.

[88] „Alain de Botton on Love," *The School of Life,* 2016. jún. 2., `https://www.youtube.com/watch?v=jJ6K_f7oSdg`; „Alain de Botton a szexről," *The School of Life,* 2012. nov. 27., `https://www.youtube.com/watch?v=osd9AKRCFRM`.

[89] De Mente, *The Korean Mind,* 335–336.

[90] Christopher Lasch, *Az önimádat társadalma*, ford. Békés Pál (Budapest: Európa Kiadó, [1979] 1984).

[91] *Artificial Intelligence: A. I. [Mesterséges értelem]* (2001) rend. Steven Spielberg.

[92] *Metropolis* (1927) rend. Fritz Lang.

[93] Milan Kundera, *Lassúság,* ford. Vargyas Zoltán (Budapest: Európa Kiadó, [1994] 1996).

egy mozifilmben vagy a színházban. A sorozatban – akárcsak a klasszikus nagyregényben – van idő a jellemek kibontakoztatására és az események és érzelmek finom ábrázolására. Míg a hagyományos dél-amerikai teleregény jellemzően a héliodóroszi *Aithiopika* képletét[94] követi, vagyis a szerelmesek többnyire már a történet elején találkoznak és egymásba szeretnek, ám különféle bonyodalmak folytán csak pár száz oldallal/epizóddal később lehetnek egymáséi – addig a koreai sorozatokban többnyire a történet közepéig várni kell, hogy ez az érzés egyáltalán megszülessen.

2018-ban két olyan új K-dráma sorozat is adásba került „a hajnali harmat országában," melyben a hős beleszeret egy robotba. Az egyik a *Nem vagyok robot,*[95] a másik az *Ember vagy te is?* című.[96] Az elsőben a Lány a robot, a másikban a Fiú. Szándékosan használom itt a nagy kezdőbetűs „a Lány" és „a Fiú" kifejezéseket. Ezek ugyanis a hagyományos magyar elnevezései a romantikus filmek fő karaktereinek; ugyanolyan megnevezések, mint például a klasszikus operában „a primadonna," „a buffo" stb.

A *Nem vagyok robot* úgy kezdődik, hogy egy üzletileg nem túl sikeres feltaláló, Jo Ji-ah, a Lány, anyagi okokból elvállalja, hogy helyettesíti a volt barátja, Hong Baek-kyun professzor és csapata által készített (és legnagyobb megdöbbenésére: róla mintázott) robotját, amikor az a legrosszabbkor vált járóképtelenné. Tulajdonosváltás miatt új kézbe került a kutatócsoport, és az új tulajdonos, az emberallergiától szenvedő Kim Min-kyu tesztelni kívánja a Santa Maria Team által kifejlesztett Aji-3 [Aji-three] nevű androidot. Mivel Kim Min-kyu betegsége nem publikus, a csapat nincs tisztában azzal, hogy egy hús-vér ember közelsége halálos veszélyt jelent a fiatalemberre. Szerencsére az meg van győződve róla, hogy nem emberrel van dolga, így pszichoszomatikus eredetű betegsége nem jelentkezik az Aji-3-nek hitt Ji-ah jelenlétében. Az alkalmi beugró azonban a tervezettnél hosszabbra nyúlik, és Jo Ji-ah, akinek el kell játszania a robotot, beleszeret a különc fiatalemberbe, az pedig beleszeret – mint gondolja – a robotba. Mivel az igazság feltárása (ti. hónapokon át becsapták) nemcsak megalázó, hanem orvosilag veszélyes is lenne, a professzor elhiteti Kim Min-kyuval, hogy törölnie kell „a robot" emlékeit. Kim Min-kyu ekkorra már a szerelemnek hála,[97] megszabadult az allergiájától, s nagy lelkesedéssel dolgozik végre emberek között. Amikor azonban a vonaton véletlenül találkozik Jo Ji-ah-val, követni kezdi a lányt, és mikor meglátja rajta az általa korábban a „robot"-nak adott nyakláncot, rájön az igazságra, és minden addiginál súlyosabb módon újra előjön az allergiája.[98]

A „robot emlékeinek törlése" után Kim Min-kyu megbánja a törlést, és tesz egy újabb kísérletet, most már az igaz Aji-3-vel, akivel (természetesen) nem érzi ugyanazt a szerelmet, mint Ji-ah-val. Bár ő úgy gondolja, ez a herakleitoszi megismételhetetlen-

[94] Héliodórosz, *Sorsüldözött szerelmesek*, ford. Szepessy Tibor (Budapest: Magyar Helikon, 1964).

[95] 로봇이 아니야 *Roboshi Aniya [Nem vagyok robot]* (2017–2018), rend. Jung Dae-yoon.

[96] 너도 인간이니 *Neodo Inganini [Ember vagy te is?]* (2018) rend. Cha Young-hoon.

[97] Boccacciótól Balassiig jól ismert motívum a szerelem jobbító, nemesítő ereje. Vö. Bán Imre, „Balassi Bálint platonizmusa," in Bán Imre, *Eszmék és stílusok* (Budapest: Akadémiai Kiadó, 1976), 122–139.

[98] Hogy a történet happy-endinggel zárul-e vagy sem, azt nem árulom el. Ahogy Kömlődi Ferenc megfogalmazta: „A koreai szerelmesfilmek […] nem úgy fejeződnek be, mint az amerikaiak…": a koreai filmek és tévédrámák nem végződnek kötelező happy-endinggel, mint a nyugati sorozatok. Nemcsak tragikus vég létezik Koreában, hanem van egy sajátosan koreai befejezés, egyfajta *keserű happy-ending* is, s azért vannak itt is boldog véget érő sorozatok. Kömlődi Ferenc, „Utak Utópiába," *Filmvilág*, 2010. dec., 18–22., 21.

 DOI 10.31400/dh-hun.2019.2.387

ség[99] miatt van így, rajta kívül mindenki más tudja, hogy nem ez a helyzet, de senki sem bolond ezt elmondani neki. Míg tehát ő azt hiszi, hogy egy robotba szerelmes, voltaképpen egy bizarr disszimuláció[100] foglya lesz, amikor Ji-ah eltitkolja magáról, hogy ember. Disszimuláció, mert úgy tesz, mintha nem lenne az, ami; és bizarr, hiszen a mesterséges intelligenciát éppen azért alkotják, hogy úgy tegyen, mintha ember lenne.

Kim Min-kyu először azt hiszi, hogy egy robotra van szüksége, aki szolgálja őt. Valójában mégsem egy robotba szeret bele – ezért sem sikerül később megismételnie, reprodukálnia a valódi robotlánnyal a közte és Ji-ah közt kialakult kapcsolatot (nem a robot, hanem saját szemszögéből). Éppen azért szereti meg a robotnak hitt lányt, mert az eredeti, kreatív egyéniség, akinek saját véleménye van. Ji-ah pedig előbb a fiatalember iránti ellenszenvétől szabadul meg, amikor megismeri Kim Min-kyu történetét, és meglátja benne a magányos kisfiút, akinek teljesen egyedül kellett boldogulnia a szülei halála után; szeretni pedig azért kezdi, mert a fiatalember zseninek tartja az ismeretlen feltalálót, aki valójában Jo Ji-ah.

Az igazság felfedése viszont több okból is szégyen lenne mindenki számára: szégyen, hogy a csapat aljas módon félrevezette a fiatalembert, és szégyen, hogy vele mindez megtörténhetett. A valódi dráma tehát nem a cselekménynek azon a pontján történik, amikor Kim Min-kyu ráébred, hogy beteges vonzalom támadt benne egy szimulakrum iránt, hanem akkor, amikor mind neki, mind a többieknek szembesülniük kell a szégyennel.

Érdemes szót ejteni a mesteri történetvezetésről is, mely először humoros részeken keresztül szeretteti meg a nézővel a karaktereket, hogy aztán komolyabbra váltva mindannyiukkal megjárassa az érzelmek poklát.

A pazar látványvilágú *Ember vagy te is?* történetében Kang So-bong, a talpraesett testőrlány nem egy magát robotnak kiadó fiatalemberbe szeret bele, hanem ténylegesen az Oh Laura[101] professzorasszony által készített Nam Shin III nevű androidba. Érdekes, hogy ez a robot is hármas számú robot, mint a *Nem vagyok robot* robotlánya, Aji-3.

Oh Laura a saját gyermekét szeretné helyettesíteni a robottal, mert kisfiát, Nam Shint elszakították tőle. Amikor újra találkozhat már felnőtt (és meglehetősen ellenszenvesen viselkedő) fiával, ahogy Nam Shin III nevezi: az *ingan* [„ember"] Nam Shinnel, úgy dönt, hogy megsemmisíti a robotot, mert nincs már többé szüksége rá.

Nam Shin III valóban a tökéletes szimulakrum: olyan, amilyennek minden anya látja a fiát: szép, kedves, jólnevelt és igazi hős. Az *ingan* Nam Shin pedig egy narcisztikus *csebol,*[102] aki gyakorlatilag mindenkit gyűlöl. A karakterek bizonyos mértékig hasonlítanak az *A. I.* főszereplőire: a robotgyermekre, a gyerek után sóvárgó anyára, illetve a szép robotfiúra. A történet pedig kimondatlanul is a „születés vagy nevelés"

[99] Közismert formában: „Nem léphetünk kétszer ugyanabba a folyóba." Platón, *Kratülosz* 402 A.

[100] A disszimuláció fogalmáról részletesen: Bókay Antal, *Bevezetés az irodalomtudományba* (Budapest: Osiris, 2006), `https://www.tankonyvtar.hu/hu/tartalom/tamop425/2011_0001_520_bevezetes_az_irodalomtudomanyba/`.

[101] 오로라 (Oh Ro-ra): A név kiejtése hasonló ahhoz, ahogy az angolban az *Aurora* nevet ejtik.

[102] 재벌 (nemzetközi átírással: *chaebeol*): koreai, családi óriásvállalat, illetve ennek örököse. Ilyen cég pl. a Daewoo, a Hyundai, a KIA, az LG és a Samsung is. Lásd még: De Mente, *The Korean Mind,* 26–29.

klasszikus kérdését idézi fel a nézőben: vajon azért „jobb ember" a robot, mert jónak programozták, vagy azért, mert jobban bántak vele, mint az emberrel?

A történet egyik pontján Nam Shin III udvariasan azt mondja, sajnálja, hogy neki nincsenek érzései, mire Kang So-bong azt feleli, hogy az ő érzései a tettekben mutatkoznak meg. Ez a mondat gyakorlatilag J. K. Rowling híres szállóigéjének parafrázisa a Harry Potter-sorozat második könyvéből.[103] Ennek a pontnak azért van jelentősége a történetben, mert segít azt úgy értelmezni, mint metaforát. *Kang So-bong mi vagyunk*: a rajongó, aki a szimulakrumba szeret bele; az imádott sztár maga pedig, mivel nem képes közvetlenül viszonozni a rajongók érzéseit, úgy viszonozza ezt, hogy újabb és újabb művekkel, szereplésekkel és fotókkal erősíti meg a szerelem retorikájával jellemzett rajongást.

Mindkét történetben különbözik a gép (robot) és az ember, akiről mintázták. Míg Aji-3 megmarad gépnek, nem lehet beleszeretni, mert nincs egyénisége, nincs személyisége, addig Nam Shin III önálló személyiséggel és önálló akarattal rendelkezik. Ezért aztán a *Nem vagyok robot* szereplői nyugodt szívvel veszik tudomásul, hogy Aji-3 csak egy állomás volt a robotikában, szétszedése, átépítése nem ér fel gyilkossággal, ezzel szemben Nam Shin III esetében gyilkosságnak számít a szereplők szemében a robot elpusztítása.

Aji-3 csak 1-es típusú android: komplex, de nem emberi rendszer. Nam Shin III a következő stádiumot képviseli, amikor az androidnak személyisége van, ugyan nem biológiai, nem szerves alapú lény, de ember.

Aji-3 csak a külsejét kapta Ji-ah-tól. A professzor tervei és Kim Min-kyu elgondolása szerint ő lenne a kiszámíthatatlan emberi lény „javított" változata, de kiderül, hogy egyértelműen a közelébe se ér a hús-vér embernek (Min-kyu a robotba nem is szeret bele). Nam Shin III is egy létező, de el nem érhető személy optimalizált változatának készült – ám (sajátos módon) emberibb és szerethetőbb lett, mint az igazi.

Az *Ember vagy te is?* témája más, ezért nem tette meg a következő lépést, hogy mi lett volna, ha Oh Laura nem a távollevő fia pótlására készíti el a robotot: mi lett volna, ha a történet arról szól, hogy elvesztett szerettének, meghalt férjének emlékeit és személyiségét „menti le" és „írja át" egy számítógépes programba. (Axiómaként fogva fel, hogy mindez megtehető, és garantálva a fizikai halhatatlanságot.)

Nem robottörténetet, hanem klasszikusabb szimulakrumszerelmet ábrázol a 2016-os *W*[104] című sorozat. A hősnő, Oh Yeon-joo itt egy képregényfigurába,[105] Kang Chulba szeret bele, miután csodálatos módon, a lassan öntudatra ébredő képregényhős miatt bekerül a képregény fiktív világának realitásába. Oh Yeon-joo szerelme nem egyszerűen reménytelenebb, mint Kang So-bongé. A *W* hősnője a történet egy pontján ráébred, hogy valamikor ő maga volt az, aki Kang Chul figuráját mint ideális férfiét megalkotta, ilyen módon az orvosnő és a webtoon-figura szerelme 21. századi Pygmalion-történetté válik. A kérdés pedig, hogy hús-vér valósággá lehet-e tenni a szimulakrumot.

[103] „A döntéseinkben, nem pedig a képességeinkben mutatkozik meg, hogy kik is vagyunk valójában." J. K. Rowling, *Harry Potter és a Titkok Kamrája*, ford. Tóth Tamás Boldizsár (Budapest: Animus Kiadó, [1998] 2000), 319.

[104] *W* – 더블유 [Deobeuryu] (2016), rend. Dae-yoon Jung.

[105] Úgynevezett *webtoon*ról (online képregényről) van szó a történetben.

 DOI 10.31400/dh-hun.2019.2.387

A számítógépes technológia fejlődése nemcsak azt teheti idővel lehetővé, hogy robotokat válasszunk életünk párjának és szerelmünk tárgyának, hanem abba az irányba is folynak kísérletek, hogyan lehet az agyunk tartalmát és személyiségünket magát digitális programba átírni, hogy ilyen módon biztosíthassuk a személyes halhatatlanságot, vagy legalábbis annak illúzióját.[106] Vagyis nemcsak androidot, emberszerű robotot próbál fejleszteni a tudomány, hanem a másik irányból, az ember felől indulva is eljuthatunk a kiborgokig. Nam Shin III is kiborgnak nevezi Kang So-bongot a lány lábába épített implantátum miatt. A *W* és az *Ember vagy te is?* valójában arra keresi a választ, hogy a virtuális világban és egy virtuálisan létező elmében hogyan lehet, illetve meg lehet-e egyáltalán védeni magunkat a „hackeléstől".

A technológia egyelőre szembemegy azzal az üzenettel, amin Attis, Alain de Botton és a többi ismert és ismeretlen lelki segítő, művész és alkotócsapat dolgozik, vagyis, hogy ne legyünk illúziók rabjai. Piccolomini, Cervantes, Flaubert és Jósika üzenetével szemben a jelenkor technológiája azt ígéri, hogy hamarosan képessé válunk arra, hogy a szimulakrumot, az illúziót valósággá tegyük. A humán tudományoknak kell megadni arra a választ, helyes-e ez. Van, amikor a döntés könnyű: a gyerekfigyelő kamera életet menthet, a *molka* viszont bűncselekményt valósít meg. A boldogság ígérete jól hangzik, de mi van, ha a saját vágyainkkal nem vagyunk tisztában, mint Kim Min-kyu. Mi van akkor, ha a digitális világba átírt valónkat a boldogság, fájdalommentesség és halhatatlanság ígéretével csalják tőrbe, de programmá válva éppen a szabad akaratunkat veszítjük el?

Az informatika egyelőre azt ígéri, idővel képes lesz arra, hogy minden titkos vágyunkat kielégítse, vagyis nem lesz szükségünk sem önismeretre, sem erkölcsre, sem etikára, mert a saját kis *Doktor Diagoras*-i (Stanisław Lem) dobozunkban az általunk felfogható világ urainak érezhetjük magunkat. Diagorasi-tartályunkban egy olyan Kínai Szobába kerülünk, ahol képtelenek leszünk megérteni önmagunkat és a világot is. A jó és a rossz tudása helyett az örök életet választjuk. Lassan egy olyan Bibliát olvasunk, ahol Ádám és Éva nem a tudás, hanem az örök élet fájáról evett, csak épp nem tudják – mert nem a tudás fájáról ettek.

A Design for Living in the Chinese Room: I. *Odi et amo*

While at the end of the 20th century information technology research offered an optimistic vision about the future of the cyber world, where the technology first of all, would help our access to the information and knowledge of humanity, in the early 21st century our cyber world has become Searle's *Chinese room*: we have started to lose control over the inputs and outputs, the information, the content, or sometimes even our own private sphere.

The first part of the study examines the emergence of various forms of cyberbullying through the „Jung Joon-yung scandal" in the K-pop

[106] fkomlodi, „Irány a halhatatlanság: a közeljövő ember-számítógép interfészei," *iMagazin*, 2017. nov. 18., `https://imagazin.hu/irany-halhatatlansag-kozeljovo-ember-szamitogep--interfeszei/`.

world. This section analyses different forms of *online shaming* from *cyber-stalking,* to sharing of *non-consensual pornography,* and even the *cyber-pillory.* The second part of the paper interprets some manifestations of virtual or cyber-love with the help of Baudrillard's concept of *simulacrum.* It is more than interesting how similarly the false (or elusive) love of the obsessive fans of Korean pop/TV/movie stars is constructed to that of simulacra, ie. robots and other imaginary heroes.

The virtual communication not just simply makes it easier to express our different kinds of human emotions by overcoming some of the factors such as self-reflection that make real communication difficult for us. At the same time, virtual communication simplifies what we say ("dislike", "love" etc.), but it also influences our way of thinking. We become different human beings by doing things in cyberspace that we would not do in real life. The *simulacrum problem* is interpreted through different Korean TV-drama-series, as the "I'm not a Robot," the "Are you human too" and the "W". The paper concludes with the discussion of some ethical problems concerning the *digitized human consciousness.*

Keywords:
digital privacy, social media, simulacrum, K-pop, K-drama

 DOI 10.31400/dh-hun.2019.2.387

Király Péter

Göttingen eResearch Alliance,

Gesellschaft für wissenschaftliche Datenverarbeitung mbH, Göttingen

peter.kiraly@gwdg.de

Marco Büchler

Institute of Computer Science, Georg-August-Universität, Göttingen

mbuechler@etrap.eu

A teljesség minőségjelzőként való mérése az Europeanában*

Az Europeana – a kulturális örökség európai digitális platformja – több, mint 3200[1] adatszolgáltatótól beérkező metaadatrekord gyűjteménye a rekordok jellemzőit tekintve meglehetősen heterogén. A rekordok eredeti típusa és kontextusa eltérő. Ahhoz, hogy hatékony szolgáltatásokat építhessünk rájuk, ismernünk kell az adatok erősségeit és gyengeségeit, más szóval a minőségét. A tanulmány egy olyan módszert javasol (és nyílt forráskódú implementálását), ami az adatok meghatározott szerkezeti tulajdonságait méri (teljesség, többnyelvűség, egyediség, rekordmintázat), hogy ezáltal minőségi problémákra világítson rá.

Kulcsszavak:
big data-alkalmazások, adatelemzés, adatgyűjtemény, szolgáltatásminőség, minőségkezelés, metaadat, adatintegráció

Bevezetés

> Az utóbbi 24 órában sok időt pazaroltam el, mert olyan dolgokat feltételeztem a (meta)adatokról, melyek nem bizonyultak helyesnek. Hosszú időt töltöttem a kód hibaellenőrzésével, de a kód jó volt, pusztán azt nem találta meg, ami ott sem volt. A rossz feltételezések a legnehezebben elkapható programozási hibák.[2]

* Jelen tanulmány eredetileg angolul jelent meg: Péter Király and Marco Büchler, „Measuring completeness as metadata quality metric in Europeana," in 2018 IEEE International Conference on Big Data (Piscataway: IEEE, 2019), 2711–2720. https://doi.org/10.1109/BigData.2018.8622487

1 A tanulmányban szereplő számok az Europeana 2018. októberi állapotára érvényesek.

2 Felix Rau, német nyelvész a metaadat-problémák következményeiről. 2018. október 18., https://twitter.com/fxru/status/1052838758066868224

Az aggregált metaadat-gyűjtemények funkcionalitása nem független a metaadatrekordok minőségétől. A következőkben néhány, az Europeanából,[3] az európai kulturális örökség digitális platformjából vett példával világítjuk meg a metaadatok fontosságát:

(a) Az adatbázisban van néhány ezer olyan rekord, aminek a címe „Photo" (fénykép) – illetve ennek valamely szinonimája, nyelvi változata – minden további leírás nélkül. Hogyan találja meg a felhasználó azokat a objektumokat, amelyek egy bizonyos épületet ábrázolnak, ha a leírások egyáltalán nem, vagy csak pontatlanul állnak rendelkezésre?

(b) Több olyan adatszolgáltató (data provider) található az Europeana portál „intézmény" címkéjű keresési facettájában, aminek többféle névváltozata van (pl. „Cinecittà Luce S.p.A." [372 412 rekord], „Cinecittà Luce" [2405 rekord], „LUCE" [105 rekord]). Feltételezhetjük-e, hogy a felhasználó ki tudja választani az összes releváns névalakot, amikor az adott intézményhez tartozó rekordok között szeretne keresni? Ha nem, akkor a keresés nem lesz teljes.

(c) Ha az „év" facettában nem formalizált és egységes adatok vannak, akkor nem lehet interaktív időtartam-szűkítést sem végezni. Hogy interpretáljunk olyan értékeket, mint „13436" vagy „97500000", ha alapvetően valamiféle évszámot várnánk?

(d) Vannak olyan rekordok, amelyek kizárólag műszaki azonosítókból állnak, nincsenek bennük leíró jellegű mezők (cím, létrehozó, leírás, tárgyszavak stb.). Ezek a rekordok emberi szemmel egyáltalán nem értelmezhetőek. Ezen okból kifolyólag nem is támogatják az Europeana egyetlen alapfunkcióját sem.

(e) Többnyelvű környezetben a felhasználó azt várná, hogy ha egy ismert entitásra pl. Leonardo mesterművére, a Mona Lisára különféle nyelveken keres (akár a „La Gioconda", „La Joconde" kifejezésekkel), akkor ugyanazt a találati listát kapja. Ezzel szemben a különféle nyelvi változatokkal történő keresés eltérő találati listákat eredményez, ugyanis a nyelvi változatok nem tartoznak közös entitás alá.

A kérdés, hogy hogyan döntsük el, mely rekordokat kell javítani, és melyek elég jók. A „célnak való megfelelés" („fitness for purpose") a minőségbiztosítás jól ismert szlogenje, arra a koncepcióra épül, hogy a minőséget valamilyen üzleti cél kontextusában, annak megfelelően kell meghatározni. A metaadatok minőségét vizsgálva először azt kell tisztázni, hogy miért fontosak a metaadatok. Az Europeana esetében ez meglehetősen egyértelmű: digitális objektumokhoz nyújt hozzáférési pontokat. Ha a rekord – adatelemekben megnyilvánuló – tulajdonságai nem teszik lehetővé a metaadat megtalálását, a kívánt cél nem teljesül, a felhasználó nem fér hozzá a digitális objektumhoz és nem fogja azt használni. Következésképpen amellett lehet érvelni, hogy a rekord minősége rossz. Akad ugyanakkor egy fontos probléma: az összes rekord kézi ellenőrzését, a ráfordított idő és a szükséges szakértelem mennyisége miatt még egy közepes méretű gyűjtemény sem engedheti meg magának.

[3] `http://europana.eu`

 DOI 10.31400/dh-hun.2019.2.388

Jelen tanulmány egy olyan általános módszertant és skálázható szoftvercsomagot javasol megoldásként, amit mind az Europeanában, mind más, a kulturális örökséget érintő, kis vagy nagy adatmennyiséggel rendelkező gyűjteményben lehet alkalmazni.

Háttér és alapok

Az Europeana kulturális örökséggel kapcsolatos metaadatrekordokat gyűjt és szolgáltat. Az adatbázis a tanulmány megírásának idején több, mint 58 millió rekordból áll, melyek több, mint 3200 intézményből származnak.[4] A rekordokat az Europeana Adatmodell (Europeana Data Model, a továbbiakban EDM) metaadatsémának megfelelően tárolják. Az egyes intézmények EDM-ben, vagy más metaadatszabványban küldik az adataikat. Az eredeti adatformátumok, katalogizálási szabályok, nyelvek és szótárak változatosságának köszönhetően nagy eltérések vannak az egyes rekordok minőségét tekintve, ami komolyan befolyásolja az Europeana szolgáltatásainak egyes funkcióit.

2015-ben egy Europeana különbizottság megvizsgálta a metaadat-minőség problémáját és erről közre is adott egy jelentést,[5] a bizottságnak azonban – ahogy a jelentésben írják – „nem volt elég hatásköre […] a metaadat-minőség metrikáit […] vizsgálni […] .” 2016-ban alakult egy ennél szélesebb körű Adatminőségi Bizottság (Data Quality Committee, DQC).[6] Ebben különféle területekről (például a metaadatok elméleti vizsgálata, katalogizálás, tudományos kutatás, szoftverfejlesztés) érkező szakértők gyűltek össze azzal a céllal, hogy elemezzék és felülvizsgálják a metaadatsémát, megvitassák az adatnormalizálás lehetőségeit, funkcionális követelményelemzést végezzenek, és meghatározzák az egyes funkciók megvalósítását lehetővé tevő metaadatelemeket (megválaszolva afféle kérdéseket, mint „melyek az Europeana alapvető funkciói?” és „mely metaadatelemek támogatják ezeket?”). A bizottság ezenfelül egy „problémakatalógust” is épít, amely a gyakran ismétlődő, helytelen metaadat-minták gyűjteménye (egyebek mellett ilyenek a többszörösen rögzített értékek, a cím megismétlése a leírás mezőben, gépi feldolgozásra szánt értékek, pl. azonosítók elhelyezése emberi feldolgozásra szánt adatelemekben).[7] A többnyelvűség kérdései különleges hangsúlyt kaptak, lévén az Europeana természetéből adódóan törekszik a többnyelvű adatfeldolgozásra és szolgáltatásra.

Jelen kutatást a DQC-vel együttműködve, részben azon belül folytattuk. Azt tűztük ki célul, hogy módszereket és érvényes metrikákat találjunk az Europeana metaadat minőségének mérésére, és azt támogatandó kifejlesszünk egy nyílt forráskódú metaadatminőség-mérő keretrendszert (*Metadata Quality Assurance Framework*).[8] A javasolt eszköz szándékaink szerint általános célú metaadatminőség-mérő szoftver,

[4] A számokat azt Europeana kereső API-ja alapján közöljük.

[5] Marie-Claire Dangerfield et al., *Report and recommendations from the task force on metadata quality,* Technical report. (The Hague: Europeana Foundation, 2016.) `https://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf`

[6] `https://pro.europeana.eu/project/data-quality-committee`

[7] Timothy Hill and Hugo Manguinhas, *Internal DQC Problem Patterns,* Technical report (The Hague: Europeana Foundation, 2016). `http://bit.ly/2jIXQGU`

[8] Felhasználói felület elérhetősége a cikk írásának idején (hozzáférés: 2019.12.13): `http://rnd-2.eanadev.org/europeana-qa/`, forráskód és további háttérinformáció: `http://pkiraly.github.io`.

amely adaptálható különféle metaadatsémákra (a támogatni tervezett sémák többek között a MARC[9] és a Kódolt Levéltári Leírás – Encoded Archival Description, EAD[10]).

A szoftver skálázható, vagyis fel van készítve nagy tömegű adatok elemzésére, együttműködik továbbá az *Apache Hadoop*[11] elosztott fájlrendszerével, az általános, nagy mennyiségű adatok feldolgozására tervezett *Apache Spark*-kal[12] és az *Apache Cassandra*[13] adatbáziskezelővel. A megközelítésmód egyik legfontosabb jellemzője, hogy képes az adatkurátorok számára érthető jelentéseket készíteni, akiknek általában a szoftverfejlesztők, adattudósok és statisztikusok által használt szaknyelvi kifejezések nem sokat jelentenek. A jelentések azok számára készülnek, akik az ott tárolt információt cselekvési tervvé tudják formálni. A keretrendszer modulokból épül fel: egy sémafüggetlen magkönyvtár mellett sémaspecifikus kiegészítések találhatók (és építhetők). Azzal számolunk, hogy az eszközt a metaadatminőség-mérés egyfajta folyamatos integrált munkafolyamatában (*continuous integration*) lehet majd használni.[14]

A kutatás azt a kérdést teszi fel, hogy hogyan lehet a kulturális örökség metaadatainak a minőségét a leghatékonyabban mérni. Általános feltevés, hogy a minőség fogalma túl összetett, és lehetetlen az összes aspektusát mérni – egyrészt elméleti szempontból (mivel például a jelenleg rendelkezésünkre álló nyelvfelismerési módszerek nem működnek jól a metaadatokban tipikus módon megtalálható rövid szövegek esetében), másrészt gyakorlati okokból (tekintve például a kutatás során rendelkezésre álló erőforrások korlátos voltát). A metaadatrekordok számos szerkezeti jellemzője azonban mérhető, és ezen mérések eredménye a legtöbb esetben jó közelítést ad. Az ilyen eredményeket hívhatnánk „metaadatszagnak", hasonlóan, ahogy a szoftverfejlesztés *kódszagnak* nevezi „azon felszíni jelzéseket, melyek rendszerint a rendszer mélyebb problémáival vannak összefüggésben".[15] A közelítés a gyakorlatban azt jelenti, hogy az eredmények önmagukban nem perdöntőek, azok arra hívják fel a figyelmet, hogy a metaadat-szakértőknek ezeket a pontokat érdemes alaposabban ellenőrizniük. Ez ugyanakkor azzal is jár, hogy az eszköz nem tárja fel azokat a hibákat, melyek nem szerkezeti sajátosságokhoz kötődnek.

A kutatás legfőbb célja, hogy rávilágítson a javítandó metaadatrekordokra. Ha megtudjuk, merre vannak a hibák, és tudunk prioritizálni, a hibák kijavíthatóak lesznek, a javításokat pedig megfontoltan tudjuk tervezni a hibák fontossági rangsorának

[9] *MAchine Readable Cataloging*, `https://www.loc.gov/marc/`. A keretrendszerre épülve készül egy MARC-vizsgáló szoftver ami elérhető a `https://github.com/pkiraly/metadata-qa-marc` címen. Meg kell jegyeznünk, hogy a MARC sokkal összetettebb szabvány, mint az EDM, és a szigorúbb szabályrendszer megléte sokkal fontosabbá teszi az egyedi problémák kiszűrését a MARC esetében az Europeana rekordjainál, így ott a hangsúly a „pontosság" és a „követelményeknek való megfelelés" metrikákra esik.

[10] `http://www.loc.gov/ead/`

[11] `http://hadoop.apache.org/`

[12] `http://spark.apache.org/`

[13] `http://cassandra.apache.org/`

[14] Lásd `http://pkiraly.github.io/2016/07/02/making-general/` és Péter Király, „Towards an extensible measurement of metadata quality," in *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017*, (New York: ACM Press, 2017) 111–115. `https://doi.org/10.1145/3078081.3078109`

[15] A fogalmat Kent Beck vezette be és Martin Fowler terjesztette el *Refactoring* című könyvében, lásd `https://martinfowler.com/bliki/CodeSmell.html`.

 DOI 10.31400/dh-hun.2019.2.388

megfelelően. Mivel az Europeana egy adataggregátor, a javításokat az információ forrásánál, az adott adatszolgáltató adatbázisán belül kell elvégezni. A minőségileg jobb adatok megbízhatóbb funkciókat támogatnak, így a gyenge minőségű rekordok kijavításával az Europeana erősebb szolgáltatásokat képes építeni. A tipikus hibák megtalálása másrészről az alapul szolgáló metaadatséma és annak dokumentációja fejlesztéséhez is elvezethet (bizonyos hibák feltehetően a séma dokumentációjában előforduló nyelvhasználat nem egyértelmű megfogalmazásaiból fakadnak), továbbá a mérés során olyan rekordokat lehet találni, melyek illusztrálják egyes metaadatelemek helyes vagy helytelen használatát. Végezetül a kiemelkedő minőségű metaadatrekordok használhatók a „követendő metaadat-gyakorlatok" elterjesztésére, vagy új szolgáltatások prototípusainak elkészítésekor.

Kutatási helyzetkép

Az elmúlt évtizedben a metaaadatok minőségmérésének informatikai alapú módszerei megjelentek a kulturális örökség területén.[16] Legutóbb Palavitsinis értékelte a téma releváns eredményeit.[17] A kulturális örökség területével némileg átfedő kapcsolt adatok (Linked Data) területén alkalmazott metrikákat Amrapali Zaveri és szerzőtársai összegezték.[18] Az ezekben hivatkozott tanulmányok meghatározták a minőség metrikáit, és számítási módszereket is javasoltak. Többnyire azonban kisebb rekordhalmazokat és az EDM-nél egyszerűbb metaadatsémákat elemeztek, továbbá általában homogénebb adathalmazokra alkalmaztak módszereket (jelentősebb kivételek a 7 millió rekordot elemző Newman és munkatársai,[19] valamint a 25 millió rekordot elemző Harper). Jelen kutatás újdonsága az, hogy megnöveli az elemzett rekordok számát, új adatvizualizációs megoldásokat és minőségjelentéseket vezet be, és más gyűjteményekben is újrahasznosítható nyílt forráskódú implementációt kínál.

A kulturális örökség metaadat értékeléséről szóló bibliográfiát lásd a Zotero hivatkozáskezelő rendszer „Metadata Assessment" (metaadat-értékelés) nevű könyvtárá-

[16] Thomas R. Bruce and Diane I. Hillmann, „The continuum of metadata quality: Defining, expressing, exploiting," in D. Hillman and E. Westbrooks, eds. *Metadata in practice* (ALA Editions, 2004) 238–256., Besiki Stvilia, Les Gasser, Michael B. Twidale and Linda C. Smith, „A framework for information quality assessment," *Journal of the American Society for Information Science and Technology* 58, 12. sz. (2007): 1720–1733. https://doi.org/10.1002/asi.20652, Xavier Ochoa and Erik Duval, „Automatic evaluation of metadata quality in digital repositories," *International Journal on Digital Libraries* 10, 2. sz. (2009): 67–91. https://doi.org/10.1007/s00799-009--0054-4, Corey Harper, „Metadata Analytics, Visualization, and Optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA)," *The Code4Lib Journal* 33. sz. (2016) http://journal.code4lib.org/articles/11752

[17] Nikos Palavitsinis, *Metadata Quality Issues in Learning Repositories*. PhD thesis, (Alcala de Henares, 2014) https://www.researchgate.net/publication/260424499_Metadata-_Quality_Issues_in_Learning_Repositories

[18] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann and Sören Auer, „Quality assessment for linked data: A survey," *Semantic Web* 7, 1. sz. (2015): 63–93. https://doi.org/10.3233/SW-150175

[19] David Newman, Kat Hagedorn, Chaitanya Chemudugunta and Padhraic Smyth, „Subject metadata enrichment using statistical topic models," in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '07*, (New York: ACM, 2007) 366–375. https://doi.org/10.1145/1255175.1255248

ban,[20] amelyet az amerikai Digitális Könyvtári Szövetség (Digital Library Federation) Metaadat-értékelés csoportja[21] és a DQC tagjai, köztük jelen dolgozat első szerzője állított össze.

Módszertan

Az EDM-séma

Az EDM-rekord[22] több entitásból tevődik össze. A rekord magja az *adatszolgáltató proxyja* (*provider proxy*), ami azokat az adatokat tartalmazza, amit az egyes szervezetek (*adatszolgáltatók*) az Europeanába beküldtek. Az adatok eredeti formátuma lehet EDM vagy számos egyéb, a kulturális örökség területén használatban lévő metaadatséma (például *Dublin Core*, EAD, MARC stb.). Ez utóbbi esetben az adatszolgáltatók vagy az Europeana átalakítja ezeket EDM-re. A rekord további lényeges részei a *kontextuális entitások* (*contextual entities*): résztvevők (*agents*), fogalmak, helyek és idősávok (*timespans*) – azon entitások (személyek, helynevek stb.) leírását tartalmazzák, melyek valamiféle kapcsolatban állnak a rekord tárgyával. Ezen kontextuális entitásoknak két fontos tulajdonságuk van:

(1) A forrásuk valamilyen többnyelvű szótár, így példányaik a nevüket több nyelven rögzítik.
(2) Amennyiben lehetséges, az entitások kapcsolódnak más entitásokhoz (a kapcsolati típusokat a Simple Knowledge Organization System (SKOS) ontológia[23] definiálja).

Az utolsó itt tárgyalt entitás neve *Europeana proxy*. Szerkezetileg megegyezik az adatszolgáltató proxyjával, de ez csak az adatszolgáltató proxy elemeit kontextuális entitásokkal összekötő linkeket tartalmaz, melyeket egy automatikus szemantikus gazdagító eljárás alakít ki.

Minden adatelem egy vagy több, az adatra épülő funkcionalitást vagy szolgáltatást tesz lehetővé. Az Adatminőségi Bizottság elemzi a funkcionális követelményeket, aminek során a tipikus felhasználói forgatókönyvek alapján (t.i. hogyan lépnek kapcsolatba a gyűjteménnyel) meghatározza a legfontosabb funkciókat, és elemzi, hogy mely metaadatelemek támogatják ezeket.[24] Vegyük például a többnyelvű visszakeresést (*cross-language recall*). A bizottság által megállapított felhasználói forgatókönyv a következő: „Felhasználóként az Europeana gyűjteményeiben az általam leginkább ismert nyelven szeretnék keresni, ugyanakkor szeretnék biztos lenni abban, hogy a dokumentumok nyelvétől függetlenül a legrelevánsabb találatokat kapom." Az említett

[20] `https://zotero.org/groups/metadata_assessment`
[21] `https://dlfmetadataassessment.github.io/`
[22] Az EDM-dokumentáció, útmutatók és más anyagok megtalálhatóak a `https://pro.europeana.eu/page/edm-documentation` címen.
[23] `https://www.w3.org/2004/02/skos/`
[24] Timothy Hill, Valentine Charles and Antoine Isaac, *Discovery – User Scenarios and their Metadata Requirements – v.3,* Technical report (The Hague: Europeana, 2015) `https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_WG/DataQualityCommittee/DQC_DiscoveryUserScenarios_v3.pdf`

 DOI 10.31400/dh-hun.2019.2.388

kontextuális elemek jórészt többnyelvűek. A funkcionalitást „lehetővé tevő" (*enabling*) elemekre vonatkozó követelmény: „minden literál értékadást[25] támogató EDM elemet nyelvi címkével *kell* ellátni, ezen felül *javasolt* az EDM elemek többnyelvű kontextuális entitáshoz való kapcsolása."[26]

Mivel ezen lehetővé tevő elemek meghatározása egyelőre nincs összhangban a mérés céljával és a meglevő rekordok tulajdonságaival, egy *aldimenziók*nak nevezett egyszerűbb modellel kezdtünk dolgozni. Ennek a modellnek az alapja az összetettebb felhasználói forgatókönyvek helyett a bizottság két tagja, Valentine Charles és Cecile Devarenne által alkotott mátrix, ami az általános funkciókat (aldimenziókat) kapcsolja össze az ezeket lehetővé tevő elemekkel. Az aldimenziók a következők:

- *Kötelező elemek* – azon mezők, melyeknek minden rekordban jelen kell lenniük. A modell kezelni tudja az olyan mezőcsoportokat, melyekből legalább az egyiküknek jelen kell lenni, pl. a tárgyszó típusú elemekből (dc:type, dc:subject, dc:coverage, dcterms:temporal, dcterms:spatial) legalább egynek;
- *leírhatóság* (*descriptiveness*) – mennyi információt hordoz a metaadat ahhoz, hogy leírja azt a tárgyat, amiről szól;
- *kereshetőség* (*searchability*) – a keresés során leggyakrabban használt mezők;
- *kontextusba helyezhetőség* (*contextualization*) – annak alapja, hogy csatolt entitásokat (személyek, helyek, időpontok stb.) találjunk a rekordban;
- *beazonosítás* (*identification*) – az objektum egyértelmű beazonosítását segítő mezők;
- *böngészés* (*browsing*) – az Europeana-portál böngészési jellemzői;
- *megtekintés* (*viewing*) – a portálon való megjelenítésben segítő mezők;
- *újrahasznosíthatóság* (*re-usability*) – a metaadatrekordok más rendszerekben való felhasználását lehetővé tevő mezők;
- *többnyelvűség* (*multilinguality*) – a többnyelvűség szempontjai, hogy a rekordok minden európai polgár számára érthetőek legyenek.

A tanulmány írása idején a modell csak a mezők meglétét vizsgálja, nem ellenőrzi, hogy tartalmuk megfelel-e az elvárásoknak – ezt a feladatot a kutatás egy későbbi pontján oldjuk meg.

Mérés

Minden rekord esetében számos olyan jellemzőt mérünk, melyek kapcsolatosak a rekord minőségével. A főbb tulajdonságcsoportok a következők:

- *egyszerű teljesség* (*completeness*) – a rekordban meglevő mezők aránya a sémában definiáltakhoz képest;
- *az aldimenziók teljessége* – adott funkciót támogató mezőcsoportok, lásd fentebb;

[25] Közvetlenül megadott érték, pl. szám, karaktersorozat, szemben a referenciális értékekkel, mint amilyen az EDM-ben az URL.

[26] Uo., 9–10.

- *mezők megléte és számossága* – mely mezők vannak jelen a rekordban és hányszor;
- *problémakatalógus* – ismert metaadat-problémák jelenléte;[27]
- *a leíró mezők egyedisége* (cím, egyéb cím, leírás);
- *többnyelvűség*;[28]
- *rekordminták* – mely mezők alkotják a „tipikus rekordokat".

A mérés három szinten történik: az egyes rekordok esetében, a gyűjtemény részhalmazaiban (pl. egy adatszolgáltató összes rekordja), végül a teljes adathalmazon.

Az első szinten az eszköz végigjár minden metaadatrekordot. Ezeket elemzi, az egyes rekordok mérési eredményeit pedig egy vesszővel határolt (*comma separated values*) fájl soraiba menti. Összességében minden rekord esetében több, mint ezer mérési eredményt (pontszámot) vagy egyéb jellemzőt nyerünk ki, melyek mindegyike egy-egy mező, mezőcsoport vagy a teljes rekord valamilyen minőséggel összefüggő tulajdonságát jelenti. Az eredményeket különféle pontozási algoritmusok számolják ki.

A második szint a részhalmazoké. Jelenleg a következő részhalmazokkal számolunk: az Europeana adathalmazokban (datasets) azok a rekordok találhatóak, melyek ugyanabban az adatgyűjtési (*data ingestion*) folyamatban kerültek be a gyűjteménybe (ezek a rekordok általában ugyanazon az átalakítási folyamaton mennek keresztül, amikor az Europeana letölti őket az adatszolgáltatóktól); az ugyanazon adatszolgáltatóktól származó rekordok; ugyanazon köztes szolgáltatóktól származó rekordok (az Europeana és az adatszolgáltatók között túlnyomó esetben van egy köztes réteg, egy olyan szolgáltató, ami tematikus vagy regionális alapon koordinál egy adatszolgáltatói csoportot); azonos nyelvű rekordok; azonos országból érkező rekordok. Az első három esetben leképezzük az alhalmazok metszetét is, amelybe azok a rekordok kerülnek, amelyek esetében mind a három vagy kettő tulajdonság közös (pl. ugyanabból a gyűjteményből és köztes szolgáltatótól származó rekordok). A jövőben a DQC kibővítheti

[27] Ez a mérés az Europeana kontextusában kísérleti fázisban van. A teljes problémakatalógust formálisan a Shapes Constraint Language (SHACL) szabvánnyal tervezzük leírni, lásd Holger Knublauch and Dimitris Kontokostas, *Shapes constraint language (SHACL),* W3C recommendation (W3C, 2017. júl. 20.) `https://www.w3.org/TR/2017/REC-shacl-20170720/`.

[28] Lásd Juliane Stiller and Péter Király, „Multilinguality of metadata. Measuring the Multilingual Degree of Europeana's Metadata," in M. Gäde et al. eds., *Everything Changes, Everything Stays the Same? Understanding Information Spaces.* Proceedings of the 15th International Symposium of Information Science (ISI 2017) Schriften zur Informationswissenschaft, (Glückstadt: Werner Hülsbusch, 2017) 164–176. `https://www.researchgate.net/publication/314879735_Multilinguality_of_Metadata_Measuring_the_Multilingual_Degree_of_Europeana's_Metadata`, Valentine Charles, Juliane Stiller, Péter Király, Werner Bailer and Nuno Freire, „Evaluating data quality in europeana: Metrics for multilinguality," in A. Caputo et al. eds., *Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries, the (Meta)-Data Quality Workshop and the Workshop on Modeling Societal Future co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017)* (Aachen: CEUR, 2017). `http://ceur-ws.org/Vol-2038/paper6.pdf`, valamint Péter Király, Juliane Stiller, Valentine Charles, Werner Bailer and Nuno Freire. "Evaluating Data Quality in Europeana: Metrics for Multilinguality," in *Metadata and Semantic Research 2018* 12th International Conference, MTSR 2018, Limassol, Cyprus, October 23–26, 2018, Revised Selected Papers (Communications in Computer and Information Science, volume 846) (Cham: Springer, 2019) 199–211. `https://doi.org/10.1007/978-3-030-14401-2_19`

 DOI 10.31400/dh-hun.2019.2.388

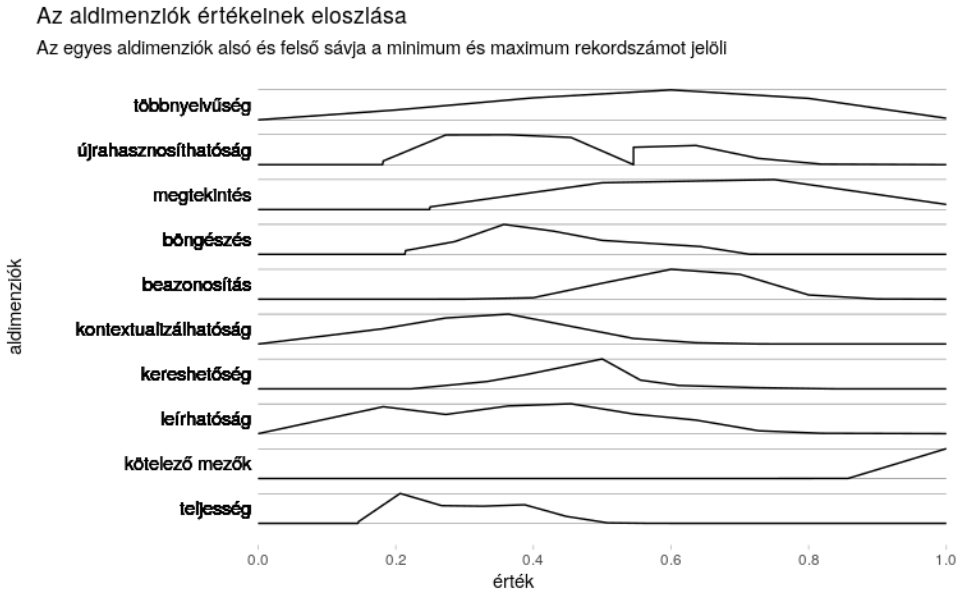
a részhalmazok körét, és olyan csoportokat is képez, melyek valamilyen másik, a metaadatsémában meghatározott tulajdonságon osztoznak.

A második és harmadik szinten az aggregált metrikákat számoljuk ki, statisztikailag összegezve az egyes rekordok esetében kiszámolt eredményeket.

A teljesség kiszámolásakor két eltérő súlyozási modellt alkalmazunk. Az első esetben a súlyok az aldimenziókat veszik figyelembe. A naiv, „egyszerű teljesség" (ahol minden adatelem ugyanazzal a súllyal szerepel) 5 súlyponttal szerepel, a kötelező elemek teljességének súlya 3, az aldimenziók súlya pedig 2. A számítási képlet (1) a következő:

$$c_{subdimensions} = \frac{\sum_{i=1}^{d} score_i \times w_i}{\sum_{i=1}^{d} w_i} \tag{1}$$

ahol d az aldimenziók száma, $score_i$ az adott aldimenzióhoz tartozó mezők közül a rekordban meglevők aránya (0-tól 1-ig terjedő skálán), a w_i pedig az aldimenzió súlya. Vagyis minden aldimenzió esetében megnézzük, hogy az ott szereplő mezők hány százalékban vannak jelen az adott rekordban, majd ezt az aldimenzió súlyának megfelelően vesszük figyelembe (ha egy aldimenzió például teljes, de a súlya alacsony, a végső pontszámnál kevesebbet fog számítani, mint egy gyengébb, de nagyobb súlyú társa). A végeredmény minimális értéke 0, maximális értéke pedig 1.



1. ábra. Az aldimenziók és az „egyszerű teljesség" értékeinek eloszlása

A második megközelítésben az elsőrendű faktor a számosság, vagyis hányszor fordul elő a mező az adott rekordban. A szélsőséges értékek torzító hatásának csökkentése érdekében nem közvetlenül ezt a számot, hanem ennek normalizált verzióját vettük alapul, amely inkább a számosság nagyságrendjét jelzi. Ilyen szélsőséges eset, amikor

egy-egy mező (pl. tárgyszó) több száz feletti példányban van jelen egy rekordban, aminek túlságosan nagy súlya lenne, s így túlságosan befolyásolná a pontszámot. A normalizálást az 1. táblázat tartalmazza.

mezőpéldányok száma	0	1	2–4	5–10	11–
normalizált pontszám	0.0	0.25	0.50	0.75	1.0

1. táblázat. A számosság normalizálása

A számosságon alapuló súlyozás egyszerű: minden mező súlya 1, kivéve az egyes entitásokat azonosító rdf:about mezőket, melyek 10 pontot kapnak, így a súlyozást főként az entitások száma és kevésbé azok „kitöltöttsége" tükrözi. Az egyenlet (2):

$$c_{cardinality} = \frac{\sum_{i=1}^{d} norm(cardinality_i) \times w_i}{\sum_{i=1}^{d} w_i} \tag{2}$$

ahol d a mezők száma, $cardinality_i$ a mezők kardinalitása, a $norm()$ a normalizálási funkció (lásd 1. táblázat) és a w_i a mező súlya. Minden mező esetében megszámoljuk, hogy hány példányuk érhető el a rekordban, a számot a táblázatnak megfelelően normalizáljuk és súlyozzuk. Eredményül – mint korábban is – egy 0-tól 1-ig terjedő számot kapunk.

A végső egyenlet a két megközelítés kombinációja, ahol az első, aldimenziós, vagyis a mezők fontosságán alapuló megközelítésmód súlya (és fontossága) két és félszer nagyobb, mint a második, az entitások és mezők számosságán alapuló megközelítésmód:

$$c = \frac{(c_{subdimensions} \times 2.5) + c_{cardinality}}{2.5} \tag{3}$$

Az alapul vett súlyszámot némileg szubjektívan állapítottuk meg, különböző mérések alapján úgy találtuk, hogy az Europeana céljainak ez az arány felel meg.

Implementáció

Az adatfeldolgozó munkafolyamatnak négy fázisa van. Az adatok forrása egy *MongoDB*-adatbázis, amiből az adatokat sororientált JSON-fájlokba exportáljuk (ahol minden sor egy külön rekord), amit Linux fájlrendszerben vagy *Apache Hadoop* fájlrendszerben tárolunk (a kutatás során rendelkezésre álló erőforrások esetében a kettő között nincs jelentős különbség, de egy több számítógépből álló klaszter esetében a *Hadoop* fájlrendszer jobb választás lehet). A rekord szintű elemzést egy, az *Apache Spark* API-t kihasználó *Java* nyelvű szoftver végzi.[29] Mivel a *Spark* automatikusan és

[29] A könyvtár magja: `https://github.com/pkiraly/metadata-qa-api`, Europeana-specifikus kiterjesztés: `https://github.com/pkiraly/europeana-qa-api`, *Spark*-felület: `https://github.com/pkiraly/europeana-qa-spark`. Az API-k (és a MARC elemzőeszköz) lefordított Java könyvtárként is elérhető a Maven központi repozitóriumban: `https://mvnrepository.com/a`

 DOI 10.31400/dh-hun.2019.2.388

konfigurálható módon támogatja a többszálú programfuttatást, az eszköz hatékonyan tudja kihasználni a futtatókörnyezet rendelkezésre álló erőforrásait (akár egyetlen, többmagos processzorú számítógépen, akár nagy kapacitású, több gépből álló számítási klaszteren dolgozunk). A számítások eredménye néhány CSV-fájl, melyeket *Apache Solr* keresőgéppel indexelünk a későbbi visszakeresés céljából – ez az eszköz kijelző felületét („műszerfal") fogja segíteni, ahol a jelentések mögött keresési találati listák húzódnak.

A harmadik fázis a rekordszintű mérési eredmények statisztikai elemzése. Ezek a szoftverek R,[30] illetve (kihasználva a Spark adatelemző API-ját) *Scala* nyelven [31] készültek. Az elemzés az előző fázisban készült CSV-fájlokat olvassa be. A kimenetet a nyers statisztikákat tartalmazó CSV- és JSON-fájlok, illetve a központi tendenciákat vagy az adatok egyéb statisztikai jellegzetességeit tükröző adatvizualizációkat tartalmazó képfájlok alkotják. Az R-nek van azonban egy gyenge pontja: kizárólag a memóriában dolgozik, így a memória mérete meghatározza a feldogozható adathalmaz méretét is. A teljes Europeana adathalmaz statisztikai elemzéséhez az általunk hozzáférhető memória kevésnek bizonyult, ezért kénytelenek voltunk a *Spark* API *Scala* nyelvű megvalósításra áttérni, és mivel a Scala statisztikai eszköztára jóval kisebb, mint a kifejezetten statisztikai elemzésekre tervezett R, ezen a ponton egyelőre kompromisszumokra kényszerültünk.

Az utolsó fázis egy online statisztikai „műszerfal", egy pehelysúlyú PHP és JavaScript alapú weboldal, ami az előző fázisok eredményeit mutatja be.[32] Az összes fázis egyetlen, közepes teljesítményű számítógépen fut (Intel Core i7-4770 Quad-Core processzor, 32 GB DDR3 RAM, Ubuntu 16.04 operációs rendszer), amit párhuzamosan más kutatás-fejlesztési projektek is használtak, ezért az, hogy a számítások erőforrás-kímélőek legyenek a szoftvertervezés során, fontos szemponttá vált.

A számítás adatforrását az Europeana-adatokról készült mentések képezik. Az első ilyen mentés 2015 végén készült az Europeana OAI-PMH szolgáltatása segítségével, amely 46 millió rekordot, 1747 adathalmazt és 3550 adatszolgáltatót tartalmaz.[33] A kutatás időtartama alatt további mentések készültek, a legutolsó 2018 augusztusában (62 millió rekord, összesen 1.27 TB-nyi fájl, az adatforrás ezúttal az Europeana *MongoDB* adatbázisának másolata volt).[34] A DQC célja, hogy havi frissítési ciklust vezessen be, vagyis az Europeana élő adatbázisa és az adatminőséget jelentő weboldal frissítése között ne legyen egy hónapnál hosszabb különbség.

`rtifact/de.gwdg.metadataqa`, így ezek más által írt Java vagy Scala szoftvercsomagokban is felhasználhatóak.

[30] Forráskód: `https://github.com/pkiraly/europeana-qa-r`

[31] `https://github.com/pkiraly/europeana-qa-spark/tree/master/scala`

[32] Forráskód: `https://github.com/pkiraly/europeana-qa-web`

[33] Az adatszolgáltatók neve nincs normalizálva, így előfordul, hogy ugyanaz az intézmény több különböző néven is szerepel.

[34] A kutatás reprodukálhatóságának kedvéért három teljes mentés elérhető a `http://hdl.handle.net/21.11101/0000-0001-781F-7` címről. Az elsőt hosszú távra archiváltuk a göttingeni Humanities Data Centre-ben: `https://hdl.handle.net/21.11101/EAEA0-826A-2D06-1569-0`. A mentések formátuma JSON, soronként egy rekorddal.

Eredmények

Teljesség

Az aldimenziókon (ahol a mezők fontossága számít) és a mezők számosságán (ahol a mező-előfordulások száma a döntő) alapuló megközelítések pontszámainak összehasonlítása rávilágít az eltérő eredményekre. Pearson-féle korrelációs koefficienssel kifejezve 0.59 a korrelációjuk, azonban eloszlásuk alakja és elhelyezkedése különbözik. A számítás módja miatt az összetett pontszám az első megközelítésmódhoz áll közelebb, a számosságon alapulónak kisebb hatása van a végső pontszámra. Az aldimenzión alapuló pontszám alsó és felső határa 0.22 és 0.92 (0–1 közötti skálán), míg a számosságon alapulóé 0.05 és 0.48. Az eloszlás részletei a 2. táblázatban és a 2. ábrán láthatók.

metrika	átlag	szórás	min.	max.
aldimenzión alapuló	0.50	0.07	0.22	0.93
számosságon alapuló	0.12	0.05	0.05	0.48
összetett	0.39	0.06	0.17	0.78

2. táblázat. A teljesség-számítás alapstatisztikái

2. ábra. A teljességszámítások eredményeinek eloszlása

Vannak olyan adatszolgáltatók, melyek összes (esetenként tízezernél is több) rekordjának ugyanaz a pontszáma; ez arra utal, hogy a rekordok struktúrája teljesen egyforma, mivel egyetlen számmal nem, csak mezőszintű elemzéssel lehet igazolni, hogy ezek a

 DOI 10.31400/dh-hun.2019.2.388

rekordok tényleg ugyanabból a (Dublin Core alapú) mezőhalmazból állnak. A másik végponton vannak azok a gyűjtemények, melyekben a pontszám nagy változatosságot mutat. Például aldimenziók tekintetében egy adatszolgáltatónak öt, 0.4-től 0.8-ig terjedő, szinte tökéletesen egyenlő eloszlású pontszáma van, míg az – ugyanezen számítás tekintetében – egyik legjobb gyűjtemény majdhogynem teljesen homogén: a rekordok 99.7%-ának pontszáma 0.9 (és még a maradék 0.3%-nak is 0.8). Ez azt jelenti, hogy az érintett mezők[35] általában nincsenek jelen az első adathalmaz esetében, de szinte mindig jelen vannak a második esetben. Az eszköz különféle ábrákat és táblázatokat kínál a pontszámok megoszlásának vizualizációjára.

A mezők eloszlásának vizsgálatából levonható első következtetés az, hogy sok rekordban nincsenek kontextuális entitások, és csak néhány adatszolgáltató rekordjaiban van 100%-os lefedettség (vagyis, ahol minden rekordban található valamelyik kontextuális entitás). A rekordok 6%-ban van *agent*, 28%-ban *place*, 32%-ban *timespan* és 40%-ban *concept* entitás. Kizárólag a kötelező technikai jellegű adatelemek érhetőek el minden rekordban. Vannak olyan mezők, melyeket ugyan definiál az adatséma, de a rekordokban egyáltalán nem szerepelnek. Vannak ugyanakkor „túlhasznált" mezők, például a dc:description-t gyakorta használják valamilyen specifikusabb mező (tartalomjegyzék, tárgyszó, egyéb cím) helyett.

A felhasználók a „műszerfalon" tudják ellenőrizni a teljes Europeana, egy adott gyűjtemény vagy egy-egy rekord minőségének jellemzőit. Az adatszolgáltatók világos képet kaphatnak az adatokról, és erre az elemzésre alapozva tervezhetik az adattisztító vagy adatjavító lépéseket.

Többnyelvűség

A DQC a közelmúltban publikálta a többnyelvűség számításának részleteit és eredményeit,[36] így ez a rész csak az eredmények rövid összefoglalása. Az EDM az RDF nyelvi annotációs modelljét követi, így az adatok létrehozói meg tudják jelölni, hogy egy adott sztring egy bizonyos nyelven íródott (például *"Brandenburg Gate"@en*, ahol a „Brandenburg Gate" a mezőérték, míg az „en" az angol nyelvet jelöli). A szerkezet neve „felcímkézett érték" (*tagged literal*). A DQC négy releváns rekord szintű metrikát azonosított.

- a felcímkézett értékek száma
- az egyedi nyelvi címkék száma
- a felcímkézett értékek száma nyelvenként
- a nyelvek átlagos száma mezőnként (azokban az esetekben, ahol legalább egy felcímkézett érték szerepel)

Ezeket az értékeket kiszámoltuk az adatszolgáltató proxyjára (vagyis az intézmények szolgáltatta eredeti adatokra), az Europeana proxyjára (ami a tipikusan többnyelvű

[35] Az adatszolgáltatói proxy dc:title, dcterms:alternative, dc:description, dc:type, dc:identifier, dcterms:created, dc:date és dcterms:issued mezői, illetve az aggregálás entitás edm:provider és edm:dataProvider mezői.

[36] Charles et al. 2018, Király et al., 2019.

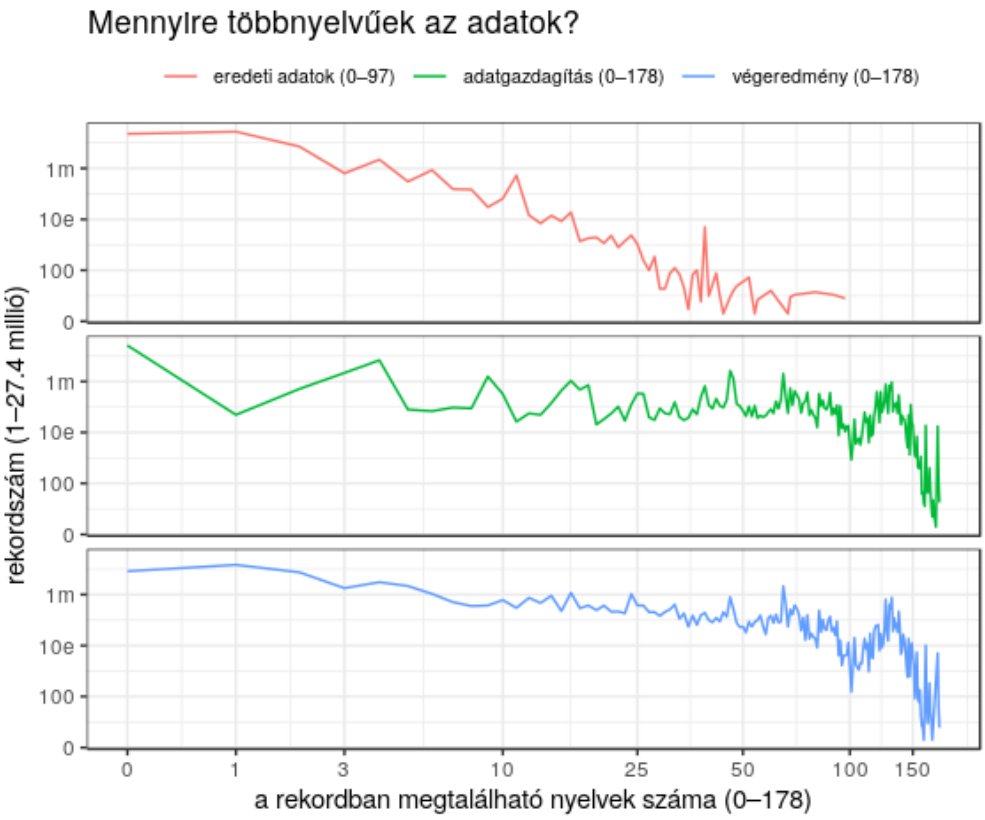
szótárakból származó adatgazdagítást tartalmazza), végül a teljes objektumra. Az eredményt a 3. és 4. táblázat és a 3. ábra tartalmazza.

metrika	szolgáltató	Europeana	teljes
a felcímkézett értékek száma	5.44	64.34	69.79
az egyedi nyelvi címkék száma	1.67	37.92	38.79
a felcímkézett értékek száma nyelvenként	2.64	0.95	2.17
a nyelvek átlagos száma mezőnként azokban az esetekben, ahol legalább egy felcímkézett érték szerepel	1.10	28.10	20.21

3. táblázat. A többnyelvűség metrikái (átlagok)

entitás	0	1	2 vagy több
szolgáltató	22.4M (36.2%)	27.3M (44.1%)	12.1M (19.6%)
Europeana	25.8M (41.7%)	49K (0.07%)	36.1M (58.2%)
teljes	8.2M (13.3%)	14.6M (23.7%)	39.1M (63.0%)

4. táblázat. A nyelvek átlagos számának eloszlása a rekordokban



3. ábra. Többnyelvűség

DOI 10.31400/dh-hun.2019.2.388

A 4. táblázat azt tükrözi, hogy csak a rekordok 20%-ában van két vagy több nyelven elérhető mező az adatszolgáltató proxyjában. Az adatgazdagítási eljárás miatt – ami többnyelvű adatforrásokból, például a DBpediából származó külső kontextuális információkat (a rekorddal kapcsolatos szereplők, fogalmak, helynevek és időpontok adatait) ad az Europeana rekordjaihoz – a többnyelvűség átfogó pontszáma magasabbá vált. Nemcsak hogy növekedett a két vagy többnyelvű mezők száma, de emellett csökkent a nyelvi annotáció nélküli rekordok száma is.

További felismerés, hogy a nyelvi címkék nem mindig szabványosak. Különféle adatszolgáltatók különféle szabványokat követnek, sőt, alkalmasint ad hoc címkéket is használnak. Az egész adathalmazban összesen több, mint 400 különféle nyelvi címke található, melyek közül több is ugyanazt a nyelvet jelöli („en", „eng", „Eng" stb. például az angolt). További kutatásokat kell a normalizált nyelvi címkékkel ellátott rekordok elemzésének szentelni, hogy helyes képet kapjunk a nyelvhasználatról.

Egyediség

A tanulmány elején említettük a hasonló címek példáját, amikor több rekordnak ugyanaz a címe. Ahhoz, hogy megtaláljuk ezeket a rekordokat, ki kell számolnunk az értékek egyediségét (*uniqueness*). Az egyediség azokban a mezőkben pozitív tulajdonság, amelyek az objektum egyedi tulajdonságait írják le, viszont kevésbé pozitív vagy egyenesen negatív azokban, amelyek a rekordokat kontextuális információkhoz kapcsolják, és ahol az értékek szükségképpen valamilyen ellenőrzött szótárból származnak, és így (ideális esetben) több rekord is ugyanazt az értéket használja. Azért, hogy hatékonyan állapíthassuk meg egy érték egyediségét, olyan keresőmotort használtunk, amelyben a mező értékeit teljes kifejezésként, az értéket elmentve indexeltünk. Mivel egy ilyen index felépítése a teljes adathalmaz összes mezőjére több erőforrást igényelt volna, mint amivel rendelkeztünk, a három, ebből a szempontból legfontosabb mezőt indexeltük: a címet (*title*), az egyéb címet (*alternative title*) és a leírást (*description*). A pontszám kiszámítására Solr relevanciaszámításának egy módosított változatát használtuk:

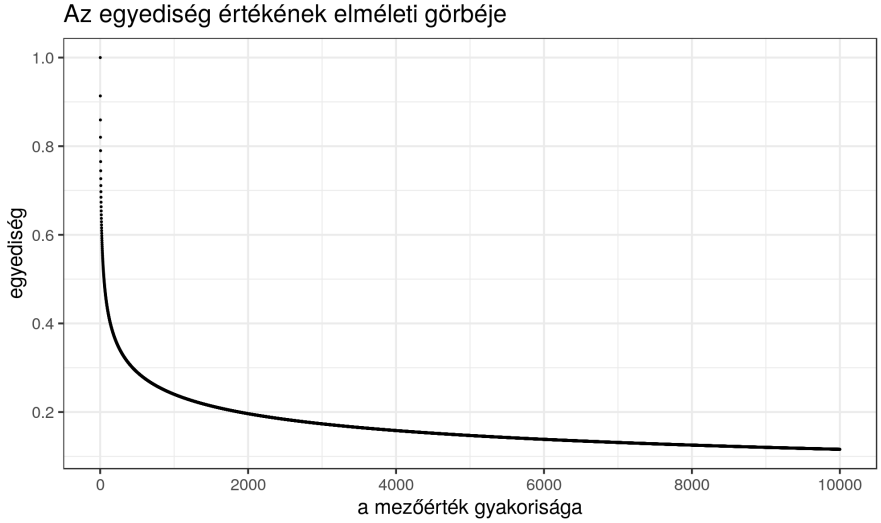
$$score(records_f,\ terms_f) = log\left(1 + \frac{records_f - terms_f + 0.5}{terms_f + 0.5}\right) \tag{4}$$

$$uniqueness_f = \left(\frac{score(records_f,\ terms_f)}{score(records_f,\ 1.0)}\right)^3 \tag{5}$$

ahol $records_f$ azon rekordok száma, amiben f mező elérhető, $terms_f$ az érték gyakorisága. A számítási mód azoknak az értékeknek ad magas pontszámot, melyek egyediek vagy nagyon kevés rekordban fordulnak elő. Minél többször fordul elő egy adott érték, az „egyedisége" annál alacsonyabb. Az eredmény minden esetben 0 és 1 közé eső szám (1 az egyedi mezőértékek pontszáma).

Ahogy a 4. ábrán látszik, a gyakoriság növekedésével a pontszám progresszíven csökken. A felhasználói felületen a következő kategorizálást vezettük be: az egyedi

értékek jelzésén túl további öt, csillaggal jelölt kategória található. A 5. táblázat jelöli a három mező kategóriahatárait.



4. ábra. Az egyediség pontszámának elméleti görbéje. Ahogy a gyakoriság nő, az egyediség pontszáma radikálisan csökken a nulla felé.

mező	*****	****	***	**	*
cím	2-	8-	37-	293-	5226-
egyéb cím	2-	6-	23-	132-	1514-
leírás	2-	7-	34-	252-	4128-

5. táblázat. A gyakoriságon alapuló egyediségkategóriák

A kategorizálás eredménye a 6. táblázatban látható. A rekordok abszolút többsége mindhárom mező esetében egyedi értékeket tartalmaz, azonban milliónyi rekordnak van egy vagy több mező esetében is alacsony pontszáma, sőt legalább tízezer rekordban a három mező közül egyik sem fordul elő. Amikor közösen vizsgáljuk a három mezőt, kiszámítva az eredmények átlagát (lásd a táblázat utolsó sorát) azt találjuk, hogy 25 millió rekord esetében mindhárom mezőnek egyedi értékei vannak, másfelől pedig a rekordoknak csak 3.62%-a tartozik a legalacsonyabb kategóriába. Ez azt jelenti, hogy bár vannak alacsony értékek, a legtöbb esetben van legalább egy mező, aminek kevésbé alacsony az értéke, vagyis nagyobb az esély rá, hogy a rekordot valamilyen keresőkifejezéssel meg lehet találni.

mező	egyedi	*****	****	***	**	*
cím	59.4	9.5	8.3	8.7	7.1	6.6
egyéb cím	62.4	11.2	7.1	3.6	2.7	12.7
leírás	54.6	9.0	7.3	10.2	6.7	11.9
közösen	45.4	10.8	15.6	18.2	6.3	3.62

6. táblázat. Mennyire egyediek az Europeana rekordjai? (%)

 DOI 10.31400/dh-hun.2019.2.388

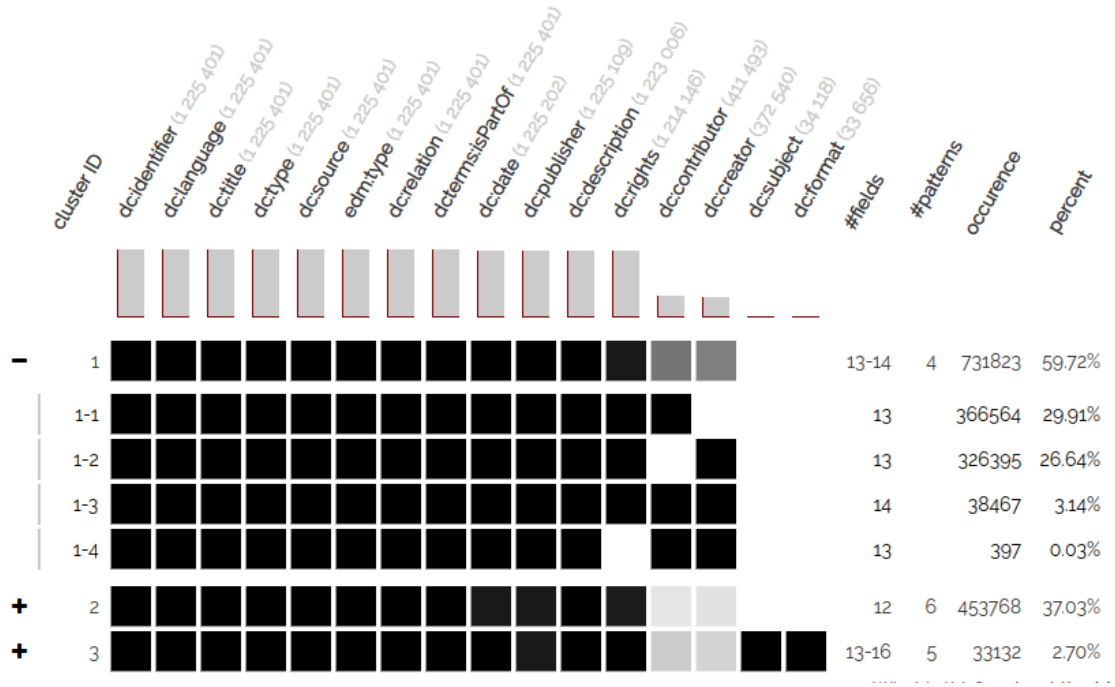
A Solr indexből ki lehet nyerni a leggyakoribb kifejezéseket. A fent említett „photograph" kifejezésen túl számos hasonló található a cím mezőben. Ezek többek között hiányzó információt (például „Unbekannt", „Onbekend" vagy „+++EMPTY+++"), gyűjtemény-, folyóirat- vagy intézménynevet („Journal des débats politiques et littéraires", „ROMAN COIN") vagy valamilyen általános leíró kifejezést („Porträtt", „Château", „Plakat", „Rijksmonument") takarnak. További vizsgálatot igényel azon gyakran előforduló kifejezések kiszűrése, melyek olyan rekordokban szerepelnek, melyekben a többi leíró mező sem rendelkezik a szükséges egyediséggel. Az eszköz megbízható kiindulópontot nyújt egy ilyen vizsgálathoz.

Rekordmintázatok

Milyen mezőkből áll a tipikus rekord? Másként fogalmazva: mely mezőket használják az adatszolgáltatók? A rekordmintázatok a rendszeresen együttálló mezők. Mivel a teljességmérés kinyeri az összes mező jelenlétét, ebből egy MapReduce algoritmus alapú elemzéssel ki lehet nyerni az együttállási mintázatokat. Itt a leképező „mapping" funkció létrehozza a mintázatot (ami a rekordban elérhető mezők rendezett listája), a redukáló „reduce" funkció pedig megszámolja azokat. Az algoritmus első megvalósításakor kiderült, hogy túl sok hasonló minta van, amelyeket érdemes lenne csoportosítani, hogy hatékonyan elemezhessük, ezért egy hasonlóan működő csoportosító algoritmust is alkalmaztunk. Ebben minden mintát először nullákból és egyesekből álló sztringgé alakítottunk. Egy-egy gyűjteményben előforduló összes mezőt szabott sorrendbe raktunk. A mezőket a következő három osztályba soroltuk: kötelező mezők, fontos mezők (melyek előfordulnak valamelyik aldimenzióban) és nem kitüntetett mezők. Ha a mező előfordul egy mintában, akkor azt egy vagy több egyes szám jelöli, máskülönben egy vagy több nulla. A kötelező mezők három számot kapnak, a fontosak kettőt, a maradék pedig egyet. Ekkor azok a minták, melyekben ugyanazok a fontos mezők, de különböznek, a nem fontos mezők közelebb kerülnek egymáshoz, mint azok, melyekben a nem fontos mezők egyeznek meg. A hasonlóságot a Jaro–Winkler algoritmussal[37] számoltuk. Megjelenítéskor (lásd 5. ábra) alapértelmezésben a csoportok jelennek meg, de a felhasználó kattintására megjelennek a csoportban szereplő egyes minták. A táblázat a csoporthoz/mintához tartozó rekordok száma alapján van rendezve, így a legtipikusabb rekordok kerültek legfelülre. Ha a mező nem érhető el minden rekordban, az azt reprezentáló négyzet szürke színben jelenik meg (a színárnyalat a rekordok számával arányos). Alapértelmezésben csak azok a csoportok jelennek meg, melyek a rekordok legalább 1%-át reprezentálják.

A rekordminták segítségével eddig kétfajta minőségi problémát tártunk fel. Az első azon rekordokra vonatkozik, melyek kevés számú mezőt tartalmaznak. Több, mint 150 000 olyan rekord van, melynek szolgáltatói proxyjában csak a következő négy mező szerepel: dc:title, dc:type, dc:rights, edm:type. Ezek közül csak az első kettő tartalmaz leíró jellegű információkat az objektumról. Nyilvánvaló, hogy a felhasználók nagy eséllyel nem lesznek képesek ezen rekordokhoz hozzáférni a facettákon keresztül, mivel hiányoznak az ehhez szükséges információk. A második probléma a szerkezeti homogeneitás: bizonyos gyűjtemények minden rekordja ugyanazokat a

[37] `https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance`



5. ábra. A rekordmintázatok csoportosítása. Az első sor a hasonló minták csoportját jelenti. A következő négy sor a csoportba tartozó egyes mintákat. A felső szürke sáv jelenti a mezők gyakoriságát az adott gyűjteményen belül.

mezőket tartalmazza. Az Europeanában összesen 906 ilyen adatszolgáltató van, de szerencsére többségük viszonylag kis gyűjtemény, csak huszonhatuknak van ezernél több rekordja. Ugyanakkor a legnagyobb homogén gyűjtemény (több, mint 500 000 rekorddal) csak 5 mezőt tartalmaz, melyből 3 leíró jellegű. Ezekkel a rekordokkal az a probléma, hogy speciális mezők helyett általános mezőket tartalmaznak (például nem tesznek különbséget a fogalmi, térbeli és időbeli tárgyszavak, osztályozások között, és eltérő típusú kontextuális információkat egyaránt a dc:type vagy a dc:subject mezőben tárolnak).

További kutatási tervek

Az Europeana a Metis[38] nevű új begyűjtési rendszer bevezetésén dolgozik, mely a tervek szerint integrálni fogja a jelen kutatás során fejlesztett eszközt. Amikor egy új rekordhalmaz importjára kerül sor, a mérés automatikusan elindul, a folyamat koordinálásával megbízott munkatárs ellenőrizheti a minőségjelentést, és ennek kimenetét, valamint saját következtetéseit megoszthatja az adatszolgáltatóval, aki reagálhat erre akár úgy, hogy megváltoztatja az adatátalakítás szabályait, akár úgy, hogy – ha lehetséges – javítják a hibákat.

A tárgyalt számítási modelleken felül számos olyan metrika van, amit kiszámítani tervezünk a közeljövőben (pontosság, információtartalom, naprakészség, ismert metaadat anti-patternek felismerése). A releváns szakirodalom azt ajánlja, hogy alakítsunk

[38] `https://github.com/europeana/metis-framework`

 DOI 10.31400/dh-hun.2019.2.388

ki egy olyan csúcs szintű pontszámot, ami az összes metrikát egy számban összegzi, így egymagában jellemzi a metaadatrekord minőségét. Ezt a metrikák súlyozásával, illetve olyan dimenziócsökkentő gépi tanuló algoritmusokkal lehet elérni, mint az elsődleges komponenselemzés (*Principal Component Analysis*).[39] Korábban említettük, hogy a jelenlegi teljességszámítási megközelítések a mező jelenlétét elemzik. Ezen a fronton a következő lépés a modell kiterjesztése a releváns mezők tartalmi értékelésének irányába, a felhasználói forgatókönyvek elemzésének megfelelően.[40]

A DQC-n belül tervezzük az itt ismertetett eredmények szakértői értékeléssel és a (naplófájlokon alapuló) használati adatokkal való összevetését. Corey Harper ismertette az Europeanához céljában és metaadatsémájában is hasonló Amerikai Digitális Közkönyvtár (Digital Public Library of America) adatain lefuttatott tesztet, amelyben azt kísérelte feltárni, hogy van-e összefüggés az objektum használata (a portálon és az API-n mérhető használati gyakoriság) és a minőségmérés során számított pontok között. A teszt sajnos sikertelen volt részben azért, mert a kutatás időpontjában még nem állt rendelkezésre statisztikai következtetések levonására alkalmas mennyiségű adat; a javasolt módszer azonban ígéretes, és ha az Europeanának elérhetőek a naplófájljai, érdemes lenne lefuttatni a kísérletet.

Tervezzük továbbá a problémakatalógus elemeinek szabatos meghatározását a W3C Shapes Constraint Language (alaki követelmények nyelve) segítségével. További tervünk, hogy az eredményeket az Adatminőségi Ontológiának megfelelő kapcsolt adatként publikáljuk.[41]

Az itt javasolt módszert más metaadatsémákra is alkalmazni lehet, például többek között MARC alapú könyvtári katalógusokra,[42] EAD alapú levéltári gyűjteményekre.[43]

Összegzés

A kutatás során (a DQC-vel együttműködésben) újragondoltuk a funkcionalitás és a metaadatséma viszonyát, és implementáltunk egy keretrendszert, amellyel sikerrel tudtuk mérni a metaadatproblémákkal korreláló szerkezeti jellegzetességeket. A keretrendszer felhasználója képes kiválogatni alacsony és magas minőségű rekordokat. Kutatási hipotézisunk szerint az olyan szerkezeti jellemzők, mint a mezők jelenléte és számossága korrelálnak a metaadat minőségével, ami igaznak bizonyult. A kutatás azáltal, hogy a korábbi szakirodalomban nem említett big data eszközöket vezetett be, kiterjesztette az elemzett rekordok mennyiségét.

A kutatás egy bizonyos adathalmazt és metaadatsémát fogott át, azonban az alkalmazott módszer általános algoritmusokon alapszik, így az más adatsémára is alkalmazható. Számos digitális bölcsészeti kutatás (néhány példa: KOLIMO [Corpus of

[39] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning*, (New York: Springer, 2013) `https://doi.org/10.1007/978-1-4614-7138-7`

[40] Hill–Charles–Isaac, *Discovery - User scenarios.*

[41] Ricardo Albertoni and Antoine Isaac, *„Data on the Web Best Practices: Data Quality Vocabulary,"* W3C note, (W3C, 2016) `https://www.w3.org/TR/2016/NOTE-vocab-dqv-20161215/`

[42] Mivel a MARC-ban számos adattartalomra vonatkozó szigorú szabály van, az EDM-ben pedig kevés, számottevő eltérés mutatkozik a két mérési módszer között.

[43] A legnagyobb európai levéltári gyűjtemény, az Archives Portal Europe (`http://www.archivesportaleurope.net/`) az adatait egy REST API segítségével CC0 licensszel publikálta.

Literary Modernism],[44] Omniart,[45] Schmidt[46]) alapul valamilyen kulturális adatbázist definiáló sémán. Ezeket a kutatási folyamatokat javítani lehet az adatforrások gyenge pontjainak feltárásával, és – Felix Raunak a dolgozat elején idézett tweetjére reagálva – a forrásokra vonatkozó pontosabb feltételezésekkel; így a következtetések is megbízhatóbbak lesznek.

Köszönetnyilvánítás

Az első szerző szeretne köszönetet mondani az Europeana Adatminőségi Bizottság régebbi és jelenlegi tagjainak; doktori kutatása témavezetőinek, Gerhard Lauernek és Ramin Yahyapournak; Jakob Voßnak, Juliane Stillernek, Mark Philippsnek a visszajezésekért; Christina Harlownak és Zaveri Amrapalinak az inspirációért; Felix Raunak a mottóért, a GWDG-nek pedig a kutatás támogatásáért.

Measuring completeness as metadata quality metric in Europeana

Europeana, the European digital platform for cultural heritage, has a heterogeneous collection of metadata records ingested from more than 3200 data providers. The original nature and context of these records were different. In order to create effective services upon them we should know the strength and weakness or in other words the quality of these data. This paper proposes a method and an open source implementation to measure some structural features of these data, such as completeness, multilinguality, uniqueness, record patterns, to reveal quality issues.

Keywords:
big data applications, data analysis, data collection, quality of service, quality management, metadata, data integration

[44] `https://kolimo.uni-goettingen.de/`

[45] Gjorgji Strezoski and Marcel Worring, „OmniArt: Multi-task Deep Learning for Artistic Data Analysis," *CoRR*, 2017. `http://arxiv.org/abs/1708.00684`

[46] Benjamin Schmidt, „Stable Random Projection: Standardized Universal Dimensionality Reduction for Library-Scale Data," in R. Lewis et al. eds., *Digital Humanities 2017. Conference Abstracts* (Montréal: McGill University–Université de Montréal, 2017) 340–342. `https://dh2017.adho.org/abstracts/497/497.pdf`

 DOI 10.31400/dh-hun.2019.2.388

<MŰHELY>

Labádi Gergely †

Szegedi Tudományegyetem, Magyar Irodalmi Tanszék

Géppel mért irodalom: a mikszáthi élőbeszédszerűség*

Bár az élőbeszédszerűség terminus a Mikszáth-szakirodalomban *A tót atyafiak* és *A jó palócok* sikere óta folyamatosan jelen van, sőt az életmű egészének fő jellegzetességévé emelkedett, a fogalmat használók nem próbálták meg részletesebben kifejteni, mit értenek rajta. A tanulmány kiindulópontja az, hogy az élőbeszédszerűségnek vannak számszerűsíthető nyelvi jellemzői, hiszen a szakirodalom elég egyértelmű és mérhető különbséget tételez a lejegyzett „irodalmi" és „beszélt" nyelv között. A dolgozat Mikszáth Kálmán és Jókai Mór egyes szövegein végzett morfológiai vizsgálatok eredményeit mutatja be.

Kulcsszavak:
regény, Mikszáth Kálmán, morfológiai elemzés, élőbeszédszerűség, szókincsgazdagság

Tahin Szabolcs 2003-ban megjelent tanulmányának címében – „»Élőbeszédszerűség« Mikszáth prózájában"[1] – nem véletlenül tette idézőjelbe az élőbeszédszerűség kifejezést. Bár a terminus a Mikszáth-szakirodalomban *A tót atyafiak* és *A jó palócok*[2] sikere óta folyamatosan jelen van, sőt az életmű egészének fő jellegzetességévé emelkedett, a fogalmat használók nem próbálták meg részletesebben kifejteni, mit értenek rajta, mindvégig megmaradt a mintha-élmény: „[…] szinte készek volnánk esküt tenni rá, hogy a szép regéket ő maga beszéli el élő szóval. […] A falu vén regélőjére gondolunk, aki téli estéken a kukoricafosztásra összegyűlt népet mulattatja mondásaival."[3]; „Mintha csakugyan eleven beszéd csengene fülembe olvasásakor…"; „Mintha nem is volna köztünk az a nagy távolság, mely írót és olvasót elválasztja…"; „Mintha minden szó egyenesen az ajkáról lebbent volna a papírosra…"; „A mese is mintha önkéntelenül buggyanna ki lelkéből…"[4] Tahin ugyanakkor nem marad a szakirodalom foglya, és nemcsak azért, mert egyértelművé teszi, Schöpflinnek és Bartának az élőbeszédszerűségből levezetett értelmezéseinek helyi értéke ma már meglehetősen

* A kutatást az EFOP-3.6.1–16–2016–00008 azonosítójú, EU társfinanszírozású projekt támogatta 2017-ben.

1 Tahin Szabolcs. „»Élőbeszédszerűség« Mikszáth prózájában," *Tiszatáj* 57, 11. sz. (2003): 53–71.

2 Mikszáth Kálmán, *A tót atyafiak* (Budapest: Grimm, 1881). Mikszáth Kálmán, *A jó palócok* (Budapest: Légrády,1882).

3 Rudnyánszky Gyula, „Mikszáth Kálmán új könyvéről," in *Mikszáth Kálmán Összes Művei* 32. köt. szerk. Bisztray Gyula, Király István (Budapest: Akadémiai Kiadó, 1968), 366–367; Tahin, „»Élőbeszédszerűség«," 54.

4 Schöpflin Aladár, *Magyar Írók* (Budapest: Nyugat, 1917), 42–43; Tahin, „»Élőbeszédszerűség«," 54.

kétes (pl. „operett íz"), hanem azért, mert feltételezi, ez az írói stratégia része lehetett. Tahin újítása, hogy a Mikszáth-szövegek elbeszélői szerepeihez kapcsolva vizsgálja az élőbeszédszerűséget. Márpedig Mikszáth szövegeiben rendszerint többféle elbeszélői hang vegyül: vannak olyanok, amelyek kifejezetten az élőbeszédhez kapcsolódnak, de vannak olyanok is, amelyek egy írásos kultúrát feltételeznek. Az elbeszélői hangok közötti különbségek pedig, mint a például felhozott *Szent Péter esernyője*[5] kapcsán Tahin igazolja, elég látványosak és meggyőzők.

Ha megnézzük a fent idézett részleteket, akkor egyértelmű, hogy a szakirodalomban oly sokszor ismételt élőbeszédszerűség az írásbeliséggel, az írásbeli kultúrával áll ellentétben. Barta János ennek irodalomtörténeti jelentőségét abban látta, hogy Mikszáth, Jókait követve lebontja a magyar regényírói hagyomány retorikusságát: „A magyar prózai epikának erős retorikus hagyományai vannak; Eötvös, Kemény írásait éppen avult retorikus jellegük teszi ma nehezen olvashatóvá. Ezt a hagyományt Jókai rendíti meg, és Mikszáth számolja fel teljesen."[6] Ez a tétel ugyanakkor a Mikszáth-értés egyik nagy keresztje – mutat rá Milbacher Róbert –, mivel egyrészt a „nagy mesélő" Jókaihoz köti Mikszáthot, másrészt pedig egy korszerűtlennek ítélt, naiv, reflexiótlan Mikszáth-próza tétele következik belőle.[7]

Kétségtelen, hogy egyes elbeszélői hangokhoz kötni az élőbeszédszerűséget termékeny stratégia, maguk a szakirodalmi idézetek is utalnak rá (pl. „a falu vén regélője"), de amint arra Tahin is felhívja a figyelmet, *A Noszty fiú…*[8] utószavában maga Mikszáth is a kötetlenül társalgó asztaltársaság fiktív befogadói szituációját ajánlja olvasóinak.

Mindazonáltal az élőbeszédszerűséget nemcsak így lehet vizsgálni. Az írásbeliség és az élőbeszéd nyelvi sajátosságait a nyelvészek elég pontosan leírták. Érsok Nikoletta összefoglalásában[9] a beszélt nyelvre a következők jellemzők: az igék magas arányban szerepelnek, de a melléknevek, jelzők aránya alacsony; a módosítószavak, kötőszavak aránya magasabb, mint írásban, gyakoriak a névmások, határozószók. Rövidebb szintaktikai szerkezetek, félbemaradt mondatok jellemzik a beszédet, gyakoriak a megszólítások, az egyszerű, egytagú mondatok, kevés a többszörösen összetett mondat, de (számomra) meglepő módon az alá- és mellérendelő összetett mondatok számarányában nem figyelhető meg lényeges különbség. Ahogy Érsok összefoglalja: „A fentiekből kifolyólag az írott nyelvi szövegek hosszabb, teljes mondatokból állnak, amelyek egymástól egyértelműen elkülöníthetők. A grammatikalitás, jólformáltság, korrektség és exaktság [sic!], továbbá a választékosabb, a normát követő szóhasználat jellemzi az írott nyelvet."[10]

A számítógépes nyelvészet ma rendelkezésre álló eszközeivel e sajátságok nagy része könnyen mérhetővé, azaz ellenőrizhetővé vált. Mindez természetesen nem azt jelenti, hogy az elbeszélői hangokhoz kötötten vizsgált élőbeszédszerűség tételét el kellene vetni, vagy, hogy így „objektívabb" elemzéseket készíthetünk. Egyszerűen an-

[5] Mikszáth Kálmán, *Szent Péter esernyője* (Budapest: Légrády, [1895]).
[6] Barta János, „Mikszáth-problémák (Első közlemény)," *Irodalomtörténeti Közlemények* 65, 2. sz. (1961): 140–161, 142; Tahin, „»Élőbeszédszerűség«," 58–59.
[7] Milbacher Róbert, „A Mikszáth-befogadás főbb irányairól," *Tiszatáj* 65, 11. sz. (2011): 80–81.
[8] Mikszáth Kálmán, *A Noszty fiú esete Tóth Marival* (Budapest: Franklin, 1908).
[9] Érsok Nikoletta Ágnes, „Szóbeliség és/vagy írásbeliség," *Magyar Nyelvőr* 130, 2. sz. (2006): 165–176.
[10] Érsok, „Szóbeliség és/vagy írásbeliség," 166.

 DOI 10.31400/dh-hun.2019.2.390

nak a lehetőségét kínálja fel a szövegek számítógépes vizsgálata, hogy a korábbiakhoz képest több és más jellegű adatra építve értelmezhessünk irodalmi jelenségeket. Jelen esetben engem az érdekel, az első két sikeres Mikszáth-kötet, *A jó palócok* és *A tót atyafiak* mérhető nyelvi sajátságai miként viszonyulnak Jókai novelláihoz, illetve saját, 1874-es elbeszéléskötetéhez, valamint, hogy egy kései példát is hozzak, az *Öreg szekér, fakó hám* novelláihoz.[11]

1. A korpusz előkészítése

A vizsgált novellakorpusz a következő: Jókaitól az 1856-os *Árnyképek* című kötet (8 novella), a *Dekameron* 1860-ban megjelent tizedik kötete (15), valamint az 1894-es *Athenaeum*-olvasótár szövegei (5) – kíváncsi voltam ugyanis egy olyan válogatásra, amely az élőbeszéd „diadalát" hozó Mikszáth-kötetek után készült.[12] Mikszáthtól az 1874-es *Elbeszélések* (8 novella), az 1881-es *A tót atyafiak* (4), az 1882-es *A jó palócok* (15) és az 1901-ben kiadott *Öreg szekér, fakó hám* (15).[13] Összesen huszonnyolc elbeszélés Jókaitól, negyvenkettő Mikszáthtól. A vizsgálatokhoz a Magyar Elektronikus Könyvtárban elérhető szövegeket választottam. A szövegek egy részét az Arcanum digitalizálta, más részét a Project Gutenberg magyar csapata. Bár textológiai szempontból jogos kritikák érhetik, de mivel egyrészt az Arcanum munkái képezik az interneten ma hozzáférhető magyar irodalmi korpusz alapját, valamint az átírás egységes szempontok alapján és egyenletes minőségben készült, valamint az Országos Széchényi Könyvtár mint befogadó intézmény mégiscsak hitelesíti, végül úgy döntöttem, kiindulásként e kísérletben érdemes elfogadni ezeket. Magam még annyit módosítottam a szükségesnek ítélt elemzés pontossága érdekében, hogy a legsúlyosabb, feltehetően a karakterfelismerés során elkövetett tévesztéseket javítottam (pl. *m* helyett *rn* – és fordítva), illetve a címeket, hogy biztosan külön elemezze őket a program, írásjellel láttam el – ha kellett. A szövegek egykorú helyesírását, ha az az értelmezést befolyásolhatta, modernizáltam, tehát például az *aszszonynyal* alakot *asszonnyal*-ra, az *a mit*, ha nyelvtanilag indokolt volt, *amit*-re javítottam.

Az elemzéshez a korpuszt a *Magyarlanc* elnevezésű nyelvi elemzővel preparáltam.[14] Az alábbi képen *A jó palócok* első novellájának *Magyarlanc*cal elemzett első néhány

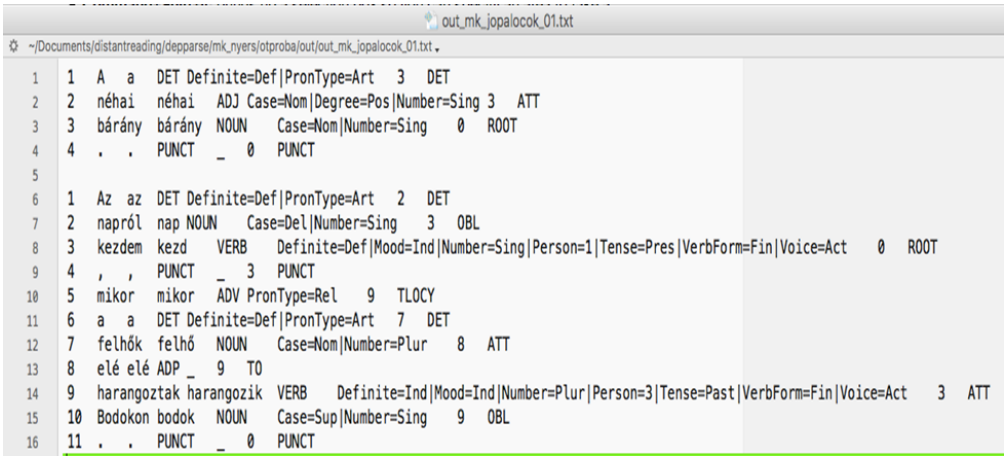
[11] Mikszáth Kálmán, *Öreg szekér, fakó hám* (Budapest: Légrády, 1901).

[12] Jókai Mór, *Árnyképek* (Pest: Emich Gusztáv, 1856), `https://mek.oszk.hu/07300/07324`; Jókai Mór, *Dekameron,* `https://mek.oszk.hu/00800/00845`, eredeti példány: *Jókai összes művei*, CD-ROM (Budapest: Arcanum, 2001); Jókai Mór, *Az egyhuszasos leány, Száz leány egy rakáson és egyéb elbeszélések* (Budapest: Athenaeum, 1894), `https://mek.oszk.hu/16300/16372`.

[13] Mikszáth Kálmán, *Elbeszélések* (Budapest: Vodiáner, 1874), `https://mek.oszk.hu/15400/15499`. Mikszáth Kálmán, *Tót atyafiak,* `https://mek.oszk.hu/00800/00895`, eredeti példány: Mikszáth Kálmán, *Gavallérok; Tót atyafiak: elbeszélések* ([Szentendre]: Interpopulart, 1995); Mikszáth Kálmán, *A jó palócok,* `https://mek.oszk.hu/00900/00950`, eredeti példány: Mikszáth Kálmán, *Tót atyafiak, A jó palócok* (Budapest: Móra, 1978); Mikszáth Kálmán, *Öreg szekér, fakó hám* (Budapest: Légrády, 1901), `https://mek.oszk.hu/11500/11563`.

[14] János Zsibrita, Veronika Vincze and Richárd Farkas, „magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian," in *International Conference on Recent Advances in Natural Language Processing,* eds. G. Angelova, K. Bontcheva and R. Mitkov (Shumen: Incoma Ltd., 2013), 763–771. *Magyarlanc,* hozzáférés: 2017.08.10, `http://www.inf.u−szeged.hu/rgai/magyarlanc`

mondata látható. A többféle elemzési lehetőség közül én a *depparse*-ot választottam, mivel ez nemcsak morfológiai elemzést végez, hanem mondattanit is.



1. ábra. *A korpusz elemzése a Magyarlanc alkalmazásával*

Az első oszlop az adott sztring mondatbeli helyét számolja, a második az eredeti szöveget tartalmazza, a harmadik a szóalakok lemmatizált, szótári alakját, a negyedik szófaját, az ötödik a szóalak morfológiai elemzését, a hatodik azt, hogy az adott szó a mondatban melyik másik szó alá van rendelve, hol van a csomópontja, a hetedik pedig a mondatbeli funkcióját. Mint a mellékelt képből látszik, a program ugyan nem mindent elemez pontosan, hiszen „bodok" valójában tulajdonnév (PROPN), nem pusztán főnév (NOUN), de azon a szinten, amire szükségem van, megfelel. Nagyon kis százalékban, de néhány, ma már nem használatos kifejezés vagy szó-, illetve ragozási alak esetében a *Magyarlanc* nem tudja értelmezni a szófajt, de ez szintén a 19. századi szövegek sajátosságából fakad, így az eredményen érdemben nem változtat.

2. Megmérni az élőbeszédszerűséget

Érsok fent idézett tanulmánya alapján elég jól meghatározhatók az élőbeszédszerűség mérhető elemei. A választékosságot a TTR, azaz a típus–jel-arány mérésével lehet számba venni: hány egyedi szóból (a képletben: V) és hány szóalakból (a képletben: N) áll a szöveg. Minél választékosabb egy szöveg, annál magasabb az egyedi szavak száma. De itt vannak még a szófaji arányok is, a mondatok hosszúságának mérése, szóval meglepően sok mindent meg lehet mérni.

3. Választékosság

Egy másik tanulmányban már foglalkoztam a szókincsgazdagság mérésének néhány aspektusával.[15] A legnagyobb probléma, hogy a szöveg hosszúságával megnő az ismétlés valószínűsége, tehát különböző hosszúságú szövegek összemérése egy egyszerű osztással, hamis eredményekhez vezet. A magyar szakirodalomban Zsilka Tibor a

[15] Labádi Gergely, „Az olvasó gép: Berzsenyi Dániel versei távolról," *Digitális Bölcsészet* 1, 1. sz. (2018): 17–34. `http://doi.org/10.31400/dh-hun.2018.1.126`.

 DOI 10.31400/dh-hun.2019.2.390

szókincsgazdaság mérésére Pierre Guiraud képletét alkalmazza – a képletet mások is elfogadják[16] –, amely, ahogy a szerző ígéri, kiküszöböli a szöveghosszból fakadó eltéréseket:

$$R = \frac{V}{\sqrt{N}}$$

Forrásai, de főleg saját vizsgálata nyomán Zsilka úgy gondolja, hogy 2000 szó alatti szövegek vizsgálatára még ez a képlet sem alkalmas. Említett cikkemben foglalkoztam e kijelentés mögött álló súlyos módszertani hibával. Meglepő módon a felvett szövegek között vannak kevesebb, mint 2000 szóalakot tartalmazó novellák. Bár teljesen véletlenszerűen választottam őket, Jókai szövegei között tizenhét, kevesebb mint 2000 szóalakból megalkotott novella található, a jóval nagyobb Mikszáth-korpuszban pedig tizenöt. Guiraud képlete így tulajdonképpen vizsgázni fog, más képletekkel is kiszámolom a szövegek „választékosságát", még ha az újabb elképzelésekkel,[17] illetve azzal, hogy érdemes-e egyáltalán efféle zárt korpuszként felfogni a szókincset, nem is foglalkozom.[18]

Ha megnézzük az alábbi grafikont, láthatjuk, hogy Jókai 1854-es kötetének folyamatosan magas szóalakszámával szemben az 1860-as kötet szinte minden írása 2000 szóalak alatti, az 1894-es viszont meglehetősen vegyes képet mutat. Utána helyezve a Mikszáth-szövegek szóhosszúságát mutató ábrát, érdekes következtetésre juthatunk. Mikszáth tulajdonképpen *A jó palócok*kal jutott el a rövidebb novellákig (a kötet alcíme: *Tizenöt apró történet*), előtte maga is közel olyan hosszúakat írt, mint a korai, korábbi Jókai-szövegek. Legalábbis ezen korpusz alapján úgy tűnik, Jókai kezdeményezte a rövidebb rövidtörténeteket a 19. század második felének magyar irodalmában. Erre a következtetésre juthatunk, ha a két korpusz egy-egy adatát szintén figyelembe vesszük. Jókai 28 novellája összesen 100302 szóalakot tartalmaz, azaz novellánként átlag 3582-t. Mikszáth jóval nagyobb korpusza 148879 szóalakot tartalmaz, ami novellánként alig valamivel kevesebb: 3545. Meglehetősen nagy a novelláskötetek által átfogott időszak, ezért a következtetés levonására a számok csak óvatosan alkalmasak: mindazonáltal utólag Mikszáth és Jókai írásművészete „egybemosódhatott" – más szóval, a Barta felvetette fejlődéstörténeti ív,[19] számszerű igazolást kaphat.

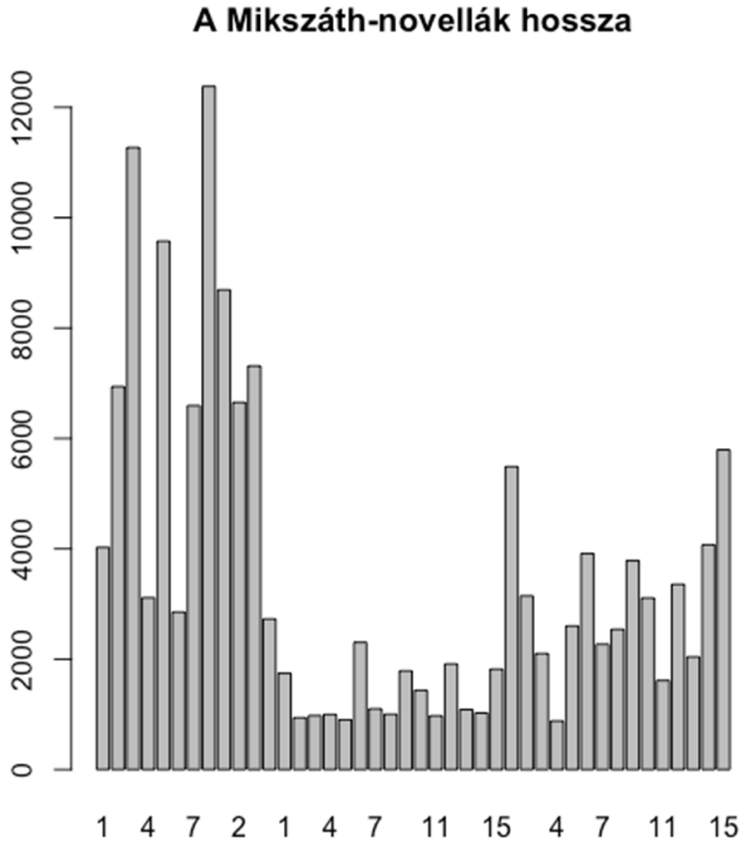
[16] Zsilka Tibor, *Stilisztika és statisztika* (Budapest: Akadémiai Kiadó, 1974). Guiraud munkájáról részletes ismertető olvasható magyarul: J. Soltész Katalin, „Guiraud statisztikai módszere a szókincs vizsgálatában," in *Általános nyelvészeti tanulmányok I.,* szerk. Telegdi Zsigmond (Budapest: Akadémiai Kiadó, 1963), 263–272.

[17] David Mitchell, „Type-token Models: a Comparative Study," *Journal of Quantitative Linguistics* 22, 1. sz. (2014), 1–21, http://doi.org/10.1080/09296174.2014.974456.

[18] András Kornai, „How many words are there?" *Glottometrics* 2, 4. sz. (2002): 61–86.

[19] Jókai és Mikszáth együtt bontják le a magyar próza retorikus hagyományát. Barta, „Mikszáth-problémák," 142.

2. ábra. *A szóalakszám változása Jókai vizsgált köteteiben*



3. ábra. *A szóalakszám változása Mikszáth vizsgált köteteiben*

 DOI 10.31400/dh-hun.2019.2.390

A szókincsgazdaság esetében persze kérdés, hány egyedi lemmából valósítja meg az író a novellát. Lejjebb látható, hogy a Guiraud-féle képlet, bár valamelyest igen, de érdemben nem csökkentette a novellák hosszúságából fakadó különbséget – Zsilkának tehát, módszertani hibája ellenére, igaza van. A Herdanhoz köthető logaritmusos képlet

$$R = \frac{logV}{logN}$$

viszont igen, amennyiben feltételezzük, hogy egy szerző stílusára a TTR többé-kevésbé jellemző – és állandó: a szókincs a szöveg hosszával kétségtelenül nő, de a függvény alapján egyre lassuló mértékben.[20]

1856	1	2	3	4	5	6	7	8
lemma	3742	1757	2397	2237	1871	3146	2034	2243
szóalak	14302	4955	7663	7900	5642	10994	6408	7098

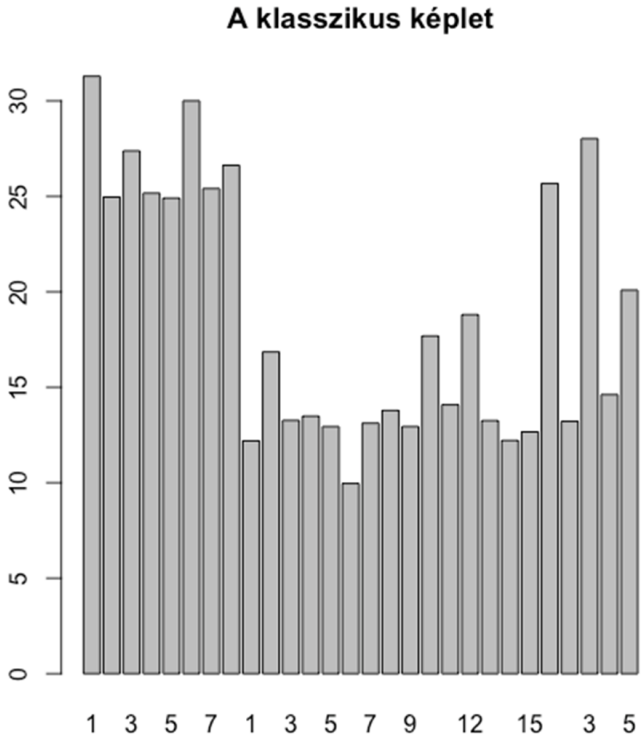
1860	1	2	3	4	5	6	7	8	9
lemma	321	564	400	352	340	179	361	406	300
szóalak	693	1120	909	681	690	323	757	867	537

1860	10	11	12	13	14	15
lemma	720	450	821	334	319	378
szóalak	1657	1021	1906	635	682	891

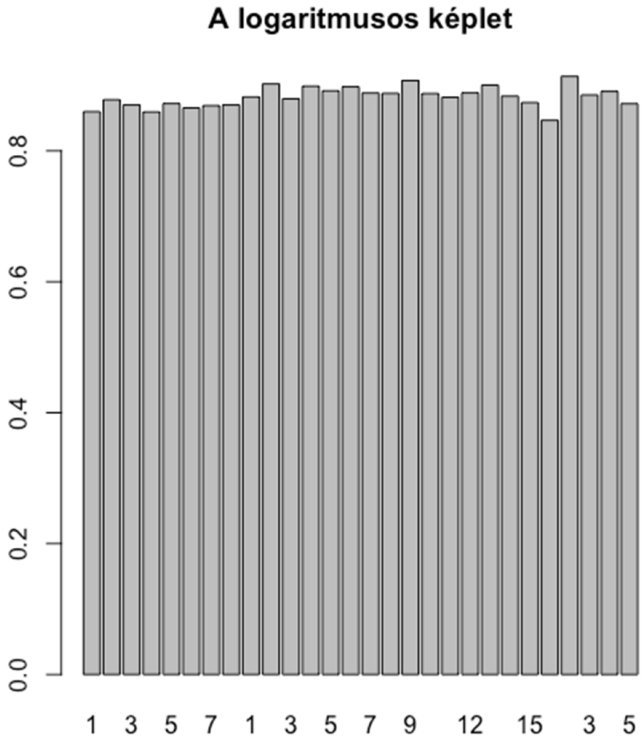
1894	1	2	3	4	5
lemma	2767	299	2116	452	1133
szóalak	11620	512	5702	956	3181

1. táblázat. *Szóalakok és egyedi lemmák száma a három Jókai-kötet novelláira lebontva*

[20] Gustav Herdan, *The Advanced Theory of Language as Choice and Chance* (Berlin: Springer-Verlag, 1966). http://doi.org/10.1007/978-3-642-88388-0.



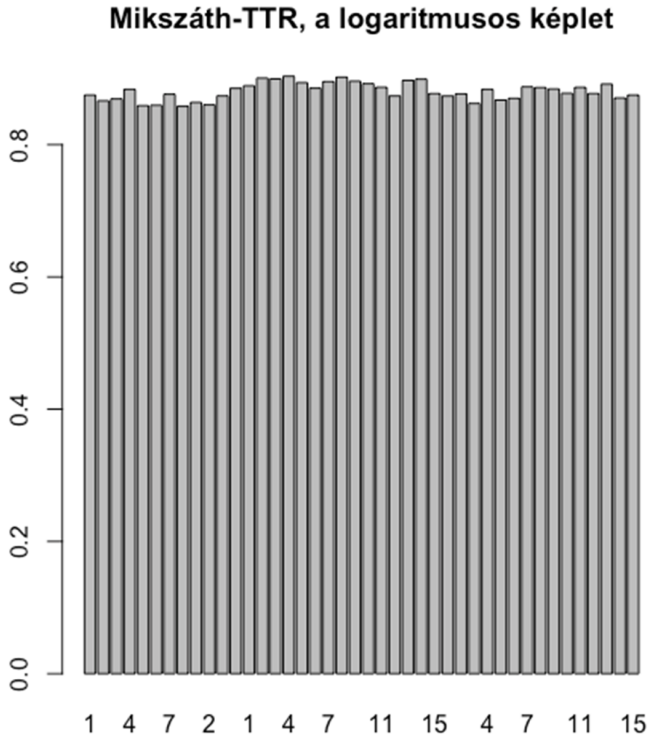
4. ábra. *Szókincsgazdagság Jókai novelláiban a Guiraud-féle képlet szerint*



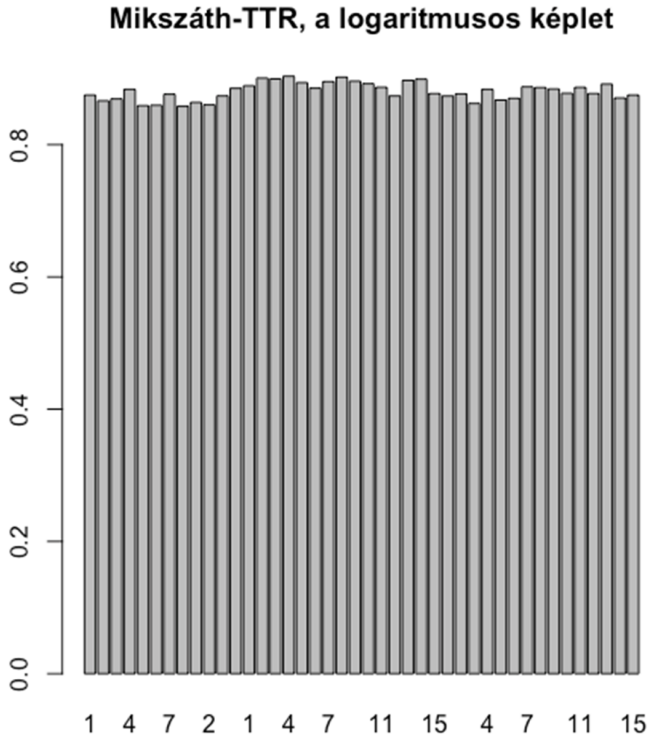
5. ábra. *Szókincsgazdagság Jókai novelláiban a Herdan-féle képlet szerint*

 DOI 10.31400/dh-hun.2019.2.390

Mikszáth esetében nem mutatom be ilyen részletesen az adatokat – a függelékben közölt *R-parancsok* alapján bárki ellenőrizheti –, de a grafikonok hasonló eredményeket mutatnak.



6. ábra. *Szókincsgazdagság Mikszáth novelláiban a Guiraud-féle képlet szerint*



7. ábra. *Szókincsgazdagság Jókai novelláiban a Herdan-féle képlet szerint*

Izgalmasabb azonban kötetekre lebontva vizsgálni az adatokat. Így – a távlati nézőpontnak köszönhetően – túlságosan homogén eredményeket kapunk. Jókai novellás-kötetre lebontott TTR-átlagai a kétféle képlet szerint, majd Mikszáthéi:

Jókai	1856	1860	1894
klasszikus	27	14	20
logaritmusos	0,87	0,89	0,88

Mikszáth	1874	1881	1882	1901
klasszikus	25	25	17	20
logaritmusos	0,87	0,87	0,89	0,88

2. táblázat. *TTR-átlagok a kétféle képlet szerint*

A logaritmusos adatok elég homogénnek mutatkoznak – már csak a kerekítések miatt is –, de ha novellákra lebontjuk a számokat, akkor azért látszanak a különbségek. Jókai szövegeinek értékei a 0,85 és 0,91 között váltakoznak, Mikszáthéi pedig a 0,86 és 0,90 között, azaz Jókaiéi valamivel változatosabbak. Ugyanakkor, ha a kiinduló kérdésre keressük a választ, azaz arra, hogy a novellák „választékossága" kifejezi-e az élőbeszédszerűséget, akkor a válasz inkább a *nem*. A mért értékek szempontjából, bár van némi különbség, ez nem tűnik jelentősnek.

4. Mérhető sajátságok

Érsok korábban idézett összefoglalása alapján azonban más nyelvi sajátosságok esetleg jobban kifejezik az élőbeszédszerűséget. Az általa említett kritériumok közül az egyik legegyszerűbben mérhető a mondathosszúság kérdése. Az alábbi táblázat Jókai novelláinak mondatszámát, illetve egy-egy novella mondatainak átlagos szószámát mutatja, a következő pedig a Mikszáth szövegeivel kapcsolatos adatokat mutatja, kötetenként átlagolva és a szélsőértékeket is feltüntetve.

mondat-/szószám	1856	1860	1894
átlag	523/17	65/13,9	341/13,8
legalacsonyabb	281/11,8	27/8,6	31/10,5
legmagasabb	930/20,1	119/20,2	998/16,5

3. táblázat. *Mondatok és szavak száma a Jókai-novellákban*

mondat-/szószám	1874	1881	1882	1901
átlag	618/11,8	506/12,9	124/10,8	273/11,6
legalacsonyabb	196/10	199/10,8	78/7,6	67/9,9
legmagasabb	989/14,6	729/15,3	196/13,6	507/13,1

4. táblázat. *Mondatok és szavak száma a Mikszáth-novellákban*

Ezek az értékek már jóval informatívabbak. Látszik, hogy Mikszáth, ha nem is feltétlenül írt rövidebb, kevesebb mondatból álló novellákat, mondatainak átlagos szóhosszúsága (értelemszerűen a tényleges szóalakokkal számoltam) már eleve jóval kisebb

 DOI 10.31400/dh-hun.2019.2.390

volt Jókaiénál. Hiába mutatnak Jókai szövegei is (illetve a belőlük készült válogatás) csökkenő tendenciát, Mikszáth maximumértékei sehol nem érik el Jókaiéit, ahogy minimumértékei is alacsonyabbak. Azaz Mikszáth szövegeire inkább igaz, hogy az élőbeszéd rövidebb, egyszerűbb mondatai dominálnak bennük.

Érsok szerint a mondathosszúság mellett a szófajok arányai is sajátosan alakulnak, a módosító- és kötőszavak gyakoribbak, ahogyan a névmások, határozószók és igék is. Szemben például a melléknevekkel. A *Magyarlanc* ún. *depparse* elemzési módja ezeket is jól kezeli. Mint fentebb a bodoki példán láttuk, az elemzés régi szövegek esetén nem mindig pontos, de mivel a nyelvi állapot miatt ugyanakkora hátránnyal indul mindkét szerző, így az eredmények értelmezhetők.

A következőkben két nagy táblázatot fogunk látni, amelyek a két szerző novellásköteteire lebontott szófaji arányait tartalmazzák. A szokásos módon először az átlag, majd a szélsőértékek is szerepelnek. A függelékben közölt *script*ek alapján természetesen részletesebb elemzések is készíthetők.

	1856	szélsőérték	1860	szélsőérték	1894	szélsőérték
ige	16,6	15,6–17,9	17,7	12,1–23,7	15,5	11,9–18,7
főnév	27,4	24,3–29,6	24,8	18,7–36,2	25,5	22,5–29,2
melléknév	10,4	8,3–11,9	8,2	5,3–11,8	10,4	7,8–13,6
kötőszó	7,4	5,9–8,7	8,5	6,5–10,6	8,1	7,4–9,1
névmás	10,6	9,1–13	12,2	9,5–18,5	10,6	8,7–13
névelő	10,1	8,4–11,4	9,5	7,3–12,7	12,3	10,8–13,2
határozó	12,7	11,4–14,7	13,7	6,8–19,3	13,2	10,2–14,9
számnév	1,1	0,8–1,5	1,5	0,6–3,7	1,6	0,8–2,9
névutó	1,8	1,4–2,3	1,5	0,6–3,4	1,3	0,5–2,1
szervetlen	0,2	0,07–0,4	0,7	0–1,9	0,3	0–0,6
igekötő	0,5	0,4–0,7	0,6	0–1,2	0,6	0,3–1

5. táblázat. *Kötetekre lebontott szófaji arányok Jókai novelláiban*

	1874	szélsőérték	1881	szélsőérték
ige	16,6	15,6–17,3	16,2	15,4–17,4
főnév	25,3	24,2–27	25,8	25,3–26,2
melléknév	10,9	9,7–12,6	10,6	9,6–11,3
kötőszó	8,3	7,4–8,9	8,7	7,7–9,1
névmás	9,5	8,5–10,5	9,2	8,6–10,1
névelő	10,7	10,2–11,2	10,4	8,8–11,4
határozó	14,4	12,9–15,4	13,9	12,9–15,3
számnév	1,2	0,8–1,7	1,3	0,9–2,2
névutó	1,2	0,9–1,4	1,2	0,9–1,6
szervetlen	0,7	0,3–0,9	0,6	0,3–1
igekötő	0,6	0,4–0,9	0,7	0,6–0,8

	1882	szélsőérték	1901	szélsőérték
ige	17,4	15,3–20,1	17,5	15,4–19,7
főnév	24,9	22,2–28,4	24,3	21,3–26,5
melléknév	9,7	7,7–12,2	9,5	7,1–10,8
kötőszó	8,1	5,3–9,8	9,2	7,8–11,4
névmás	7,5	6,2–9,8	8,3	6,4–10
névelő	12,5	10,9–14,3	13	10,7–16,1
határozó	15,4	12,9–18,1	13,7	11,6–15,7
számnév	0,9	0,5–1,8	1,1	0,3–2,1
névutó	1	0,5–1,4	1,1	0,7–1,6
szervetlen	0,9	0,4–1,3	0,8	0,2–1,2
igekötő	0,7	0,3–1,1	0,5	0,3–0,8

6. táblázat. *Kötetekre lebontott szófaji arányok Mikszáth novelláiban*

5. Következtetések

A tanulmány kiindulópontja az volt, hogy az élőbeszédszerűségnek vannak számszerűsíthető nyelvi jellemzői, a szakirodalom elég egyértelmű és mérhető különbséget tételez a lejegyzett „irodalmi” és „beszélt” nyelv között.

A felkínált sajátosságok közül, mint láttuk, a választékosság kapcsán nem sikerült olyan jellemzőt találni, amely Jókai és Mikszáth szövegei között döntő különbséget mutattak volna. A mondatok szóhosszúsága kapcsán ugyanakkor világos különbség igazolódott, még ha ennek értelmezése ennyi adat fényében nem is teljesen egyértelmű. A további mért sajátságok esetén megfigyelhettük, hogy Mikszáthnál az igék aránya lassan, de biztosan nő, amit nemcsak a kötetek átlagai, de a szélsőértékek fokozatos növekedése is mutat. Ráadásul ezzel párhuzamosan a főnevek aránya csökken. Jókai adataival összehasonlítva a különbség még feltűnőbb. Az igék aránya, ha csak kevéssel is, de Mikszáthéi alatt maradnak – ez leginkább a szélsőértékeknél mutatkozik meg –, ráadásul a főnevek, ha valamelyest csökkenő tendenciát mutatnak is, magasabbak Mikszáthéinál. Ugyanígy a kötőszavak aránya Mikszáth köteteiben magasabb, ahogyan a szervetlen közbevetéseké is. A névmások ugyan nem, de a határozószók aránya ismét igazolja a felvetést. Ha tehát az Érsok által felsorolt szófaji sajátosságokat megszámoljuk, akkor Mikszáth szövegei kétségtelenül közelebb állnak az élőbeszédhez. Ebből a szempontból éppenséggel az 1874-es kötet sikertelensége elgondolkodtató, hiszen értékei *A tót atyafiak*hoz képest nem térnek el, sőt az élőbeszédszerűség szempontjából sok esetben még „jobbak” is – persze nem állítom, hogy a siker pusztán a szövegek nyelvi sajátosságaiban rejlene.

Ugyanakkor Jókai utólag összeállított, 1894-es kötete is érdekes, mivel a szerkesztők a Jókai saját kötetei esetében már vagy három évtizeddel korábban meginduló folyamatot, – az élőbeszédhez közelítést (növekvő igearány mellett csökkenő főnévarány) – „visszafordítják”. Azaz egy Jókai-válogatást nem tudnak/akarnak – vagy éppenséggel nem lehet – úgy elvégezni, hogy a korabeli tendenciákhoz igazítsák a novellákat.

Néhány rövid történet vizsgálata önmagában persze nem elég. A későbbiekben érdemes volna egy szerzőkre, műfajokra lebontott átfogó vizsgálatot is elvégezni,

 DOI 10.31400/dh-hun.2019.2.390

mivel részletesebb, irodalomtörténetileg releváns következtetések levonásához erre volna szükség – akár a novellákat közlő lapok szintjéig eljutva, mivel a közeg sok esetben befolyásolta magát a szöveg átdolgozását is.[21]

Machine-readable Literature: "Spoken Language" in Mikszáth's Short Stories

Literary scholars have deployed the concept of "spoken language" to describe Kálmán Mikszáth's fiction since the success of his short story collections entitled *A tót atyafiak* (*Slovak Kinsmen,* 1881) and *A jó palócok* (*The Good Palots,* 1882). Although this stylistic concept has become a key characteristic feature of Mikszáth's *oeuvre,* no attempt has been made to elaborate on its definition. As scholarship assumes a clear-cut and measurable distinction between the written "literary" and "spoken" language, this paper claims that this spoken language has quantifiable linguistic markers. This is demonstrated by the morphological analysis of Kálmán Mikszáth and Mór Jókai's fictional writings.

Keywords:
fiction, Kálmán Mikszáth, spoken language, morphological analysis, lexical richness

Függelék

A függelékben a tanulmányban használt *R-script*eket adom meg, hogy bárki újrafuttathassa és elemezhesse az eredményeket, vagy saját céljaira használhassa. Én az *RStudio* nevű programot használtam. A *script*ek előtt álló számok nem a kód részei, hanem a könnyebb követhetőséget szolgálják.[22]

Fájlok behívása

Értelemszerűen a Jókai- és Mikszáth-novellákat külön dolgoztam fel, de a képletek ugyanazok, ezért a nevezéktanban nem teszek különbséget.

```
1. filenames <- list.files(path="~eleresiut", pattern="*.txt")
2. filelist <- lapply(filenames, function(x){read.csv2(x, header
   = FALSE, sep = "\t", stringsAsFactors = FALSE)})
```

[21] Török Erzsébet Zsuzsanna, „A konyhaszolgálótól Szűz Máriáig (Az irodalom hétköz- és ünnepnapi közegei a 19. század végén)" in *A Látható könyv*, szerk. Hász-Fehér Katalin (Szeged: Tiszatáj, 2006), 179–226.

[22] A kódsor a tanulmány mellékleteként TXT formátumban letölthető a cikk weboldaláról. https://doi.org/10.31400/dh—hun.2019.2.390

Mondatszám

Mivel a *depparse*-szal végzett elemzés esetében minden mondat minden elemet (szavak, írásjelek) megszámoz, ezért egész egyszerűen össze kell számolni az 1-eseket.

```
3. mondatszam.i <- sapply(filelist, function(df){dim(subset(df,
   V1==1))[1]})
```

Egyedi lemmák száma

Depparse esetén a harmadik oszlop a lemmaoszlop. Írásjeltelenítjük, listátlanítjuk. Kivesszük a ”c”-t, ami a listaformátum miatt kerül bele mindenhová az első helyre. Kiszámoltatjuk, mennyi lemma van az egyes szövegekben. Az integeres verzióval egyszerűbb számolni.

```
4. lemmak <- lapply(filelist, function(df){df["V3"]})
5. csaklemmak <- lapply(lemmak,
   function(x){strsplit(as.character(x), "(\\W+)")})
6. csaklemmak <- lapply(csaklemmak, function(x){unlist(x)})
7. csaklemmak <- lapply(1:length(csaklemmak),
   function(x)csaklemmak[[x]][csaklemmak[[x]]!="c"])
8a. egyedi.szoszam <- sapply(1:length(csaklemmak), function(x)
   unique(unlist(csaklemmak[[x]])))
8b. egyedi.szoszam.i <- sapply(1:length(egyedi.szoszam),
   function(x) length(egyedi.szoszam[[x]]))
```

Betűszám

Ezt értelemszerűen a tényleges szavakból, szóalakokból kell számolni.

```
9. szavak = lapply(filelist, function(df){df["V2"]})
10. szoalakok <- lapply(szavak,
   function(x){strsplit(as.character(x), "(\\W+)")})
11. szoalakok <- lapply(szoalakok, function(x){unlist(x)})
12. szoalakok <- lapply(1:length(szoalakok),
   function(x)szoalakok[[x]][szoalakok[[x]]!= "c"])
13. szoalakok.i <- sapply(1:length(szoalakok),
   function(x)length(szoalakok[[x]]))
14. betuszam.i <- sapply(1:length(szoalakok),
   function(x)nchar(szoalakok[[x]]))
```

Átlagos szóhossz

Hány betűből áll egy szó átlagban.

```
15. szohossz <- sapply(1:length(csakszavak), function(x)
   mean(nchar(szoalakok[[x]])))
```

 DOI 10.31400/dh-hun.2019.2.390

Átlagos mondathossz

Hány szóból áll, hány betűből áll egy mondat.

```
16. mondathossz <- egyedi.szoszam.i/mondatszam.i
17. mondathossz2 <- mondathossz*szohossz
```

Igék száma

```
18. igeszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="VERB"))[1]})
```

Főnevek száma

A tulajdonneveket is ideszámoltam.

```
19. fonevszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="NOUN"))[1]})
20. propnszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="PROPN"))[1]})
21. fonevszam.i <- fonevszam.i + propnszam.i
```

Melléknevek száma

```
22. melleknevszam.i <- sapply(filelist,
    function(df){dim(subset(df, V4=="ADJ"))[1]})
```

Kötőszószám

```
23. kotoszoszam1.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="CONJ"))[1]})
24. kotoszoszam2.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="SCONJ"))[1]})
25. kotoszoszam.i <- kotoszoszam1.i + kotoszoszam2.i
```

Névmások

```
26. nevmasszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="PRON"))[1]})
```

Névelők

```
27. neveloszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="DET"))[1]})
```

Határozószók

```
28. hatarozoszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="ADV"))[1]})
```

Számnév

```
29. szamnevszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="NUM"))[1]})
```

Névutók

```
30. nevutoszam <- sapply(filelist, function(df){dim(subset(df,
    V4=="ADP"))[1]})
```

Szervetlen közbevetések

```
31. szervetlenszavak <- sapply(filelist,
    function(df){dim(subset(df, V4=="INTJ"))[1]})
```

Igekötők

```
32. igekotok <- sapply(filelist, function(df){dim(subset(df,
    V4=="PART"))[1]})
```

Szófajok

Egy közös táblázatban összegezzük az eddigi eredményeket. Amint a tanulmányban jeleztem, a szófajok esetében tízes nagyságrendben nem tud megbirkózni a *Magyarlanc* egyes szóalakkal, ami a 100000 fölötti szóalakszámot tekintve elhanyagolható különbség.

```
33. szofajok <- data.frame(igeszam.i, fonevszam.i,
    melleknevszam.i, kotoszoszam.i, nevmasszam.i, neveloszam.i,
    hatarozoszam.i, szamnevszam.i, nevutoszam, szervetlenszavak,
    igekotok)
```

Szófaji arányok

```
34. igearany <- (igeszam.i * 100)/szoalakok.i
35. fonevarany <- (fonevszam.i * 100)/szoalakok.i
36. melleknevarany <- (melleknevszam.i * 100)/szoalakok.i
37. nevmasarany <- (nevmasszam.i * 100)/szoalakok.i
38. neveloarany <- (neveloszam.i * 100)/szoalakok.i
39. hatarozoarany <- (hatarozoszam.i * 100)/szoalakok.i
40. szamnevarany <- (szamnevszam.i * 100)/szoalakok.i
41. nevutoarany <- (nevutoszam * 100)/szoalakok.i
42. szervetlenarany <- (szervetlenszavak * 100)/szoalakok.i
43. igekotoarany <- (igekotok * 100)/szoalakok.i
44. szofajarany <- data.frame(igearany,fonevarany,melleknevarany,
    kotoszoarany, nevmasarany, neveloarany, hatarozoarany,
    szamnevarany, nevutoarany, szervetlenarany, igekotoarany)
```

 DOI 10.31400/dh-hun.2019.2.390

Választékosság

Két képletet használ a dolgozat. Hány egyedi lemma és hány szóalak alkotja a novellákat.

```
45. ttr <- egyedi.szoszam.i/sqrt(szoalakok.i)
46. ttr2 <- log(egyedi.szoszam.i)/log(szoalakok.i)
```

Szófaji arányok átlaga, szélsőértéke

A szófaji arányokat már kiszámoltuk és egy táblázatban elmentettük (szofajaranyok). Ha az egyes novellák értékeire kíváncsiak vagyunk, akkor úgy kell lekérdezni. Mivel az oszlopok az egyes szófajokat tartalmazzák, a sorok pedig az egyes novellák, csak tudni kell, melyik kötetben hány novella van. Jókainál 8, 15, 5, Mikszáthnál 8, 4, 15, 15. A szögletes zárójelben először a sorokat adjuk meg, aztán az oszlopot. Átírva értelemszerűen folyamatosan megy a sorok számozása. Tehát a 47. kód Jókai 1856-ös kötetének novelláira, azok igearányára kérdez rá. Először az átlagra, aztán a szélsőértékekre. Ezek olvashatók a dolgozatban. A 49–50. az 1860-as kötet főnévarányaira kérdez, az 51–52. pedig az 1894-es kötet mellékneveire. A Mikszáth-kötetek esetében hasonló logika alapján kell a sorokat „kiosztani”.

```
47. mean(szofajarany[1:8,1])
48. range(szofajarany[1:8,1])
49. mean(szofajarany[9:23,2])
50. range(szofajarany[9:23,2])
51. mean(szofajarany[24:28,3])
52. range(szofajarany[24:28,3])
```

<KRITIKA>

Sárhegyi Tamás Felicián
Pázmány Péter Katolikus Egyetem

sarhegyi.tamas@gmail.com

David M. Berry and Anders Fagerjord. *Digital Humanities: Knowledge and Critique in a Digital Age.* Cambridge: Polity Press, 2017.

ISBN 9780745697697

A digitális bölcsészet *(digital humanities)* keretein belül alkotott teoretikus szövegek gyakran az öndefiníciót emelik központi kérdésükké és próbálnak meg állást foglalni abban, hogy a digitális bölcsészet ténylegesen tekinthető-e önálló tudománynak vagy csupán a kor technikai vívmányait implementáló, digitális eszközök szakszerű alkalmazásának módszere.[1] A könyv is részben ilyen meta-szöveg, amely a digitális bölcsészet, valamint a kommunikáció- és médiatudománnyal összefüggő elméleti és empirikus kutatásokra épül, kifejezetten azért, hogy mélyebben meg lehessen érteni a „kritikai digitális bölcsészet" *(critical digital humanities)* fogalmát. A legfontosabb kérdés korunkban – amelyet *poszt-digitális*nak neveznek – a szerzők szerint nem az, hogy valami digitális-e avagy sem. Sokkal inkább az, hogyha valami nem digitális, mindinkább másodlagossá válik, ahogyan a kultúra különböző formái egyre inkább digitálisan termeltek, megosztottak, hozzáférhetőek és fogyasztottak lesznek. A könyv amellett érvel, hogy a digitális bölcsészetnek olyan elméleti és módszertani megoldásokat kell ajánlania, amelyek áthidalják a humán tudományok és a technika különválasztott világait. Amellett, hogy a szöveg alaposan elemzi a tudományterületet, példaszerűen alkalmazza a kritikai megközelítést.

Gyakori sztereotípia, hogy a technológia egy kész befejezett dolog a bölcsészettudományok számára, amely saját világán kívülről kell, hogy használatba vegyen a saját létének megőrzése érdekében, hiszen folyamatos támadásoknak van kitéve a 21. században célját, értelmét és relevanciáját tekintve. Következésképpen a szerzők amellett érvelnek, hogy a humán tudományoknak elméleti alapon kell megágyazni az informatika jelenségének a kultúrában, különben egyre inkább eltávolítják magukat a társadalomtól, amely feltétel nélküli bizalommal használja a digitális technológiát. A könyv fejezetei megpróbálják összefoglalni a digitális bölcsészet irányvonalait, leágazásait, és bemutatni azok létező és jövőbeni pozitív és negatív aspektusait. A szerzők emellett egy lehetséges és szükséges kritikai fordulat („critical turn") mellett érvelnek, amely megerősíthetné a tudományág státuszát az akadémiai világban. Felvázolják, hogy az egyre inkább piaci alapon működő felsőoktatási intézmények milyen

[1] A digitális bölcsészetről alkotott meta-szövegek: Ray Siemens and Susan Schreibman, eds., *A Companion to Digital Literary Studies,* Blackwell Companions to Literature and Culture (New York: Wiley–Blackwell, 2013); Steven E. Jones, *The Emergence of the Digital Humanities* (New York: Routledge, 2014); Susan Schreibman, Raymond George Siemens and John Unsworth, eds., *A New Companion to Digital Humanities* (Chichester: John Wiley & Sons Inc, 2016); James Smithies, *The Digital Humanities and the Digital Modern* (London: Palgrave Macmillan, 2017).

kapcsolatban vannak az online globális tudáspiaccal (MOOC-portálok, pl. *Coursera, TedX, Udacity*), és ez hogyan befolyásolja a digitális bölcsészet szerepét és megítélését.

A digitális bölcsészeten belül elterjedt szemlélet, hogy többet kell csinálni, mint beszélni („more hack, less yack"). Ez a megközelítés rávilágít a tudományág már korábban említett elméleti szövegeinek öndefiníciós kényszeréből fakadó egyhangúságára, amelynek ellenpontja a túlzott technikai orientáció (technofília), akár egy-egy projekten vagy képzési struktúrán belül is. Ezen másik véglet célja valami létrehozása, termelése és felépítése a teoretizálás helyett, így például digitális rendszerek létrehozása, archívum- és adatbázis-építés, digitalizálás, *mapping*. Ezen projekteken belül azonban sokszor maga a technológia kerül a fókuszba. A humán tudományok iránti csekély érdeklődést néha megtöri a nagyközönség számára is vonzóbb(nak tűnő) digitális bölcsészeti projektek jelenléte. A szerzők a túlzottan technológia fókuszú és a szélsőségesen teoretizáló kutatások között látják a kiutat, hogy a digitális bölcsészet megtalálja a helyét. A digitális bölcsészet képzések és kutatási projektek létrehozása a szerzők szerint van, hogy csak arra irányulnak, hogy egyetemek, kutatóközpontok hagyományos képzési programjaikat felfrissítsék és vonzóbbá tegyék, vagy egyéb forrásokat vonjanak be a működésükbe. A felvázolt dilemmák a digitális bölcsészettel kapcsolatban a szerzők szerint nem függetleníthetők a társadalomban végbemenő változásoktól (digitalizáció és digitális eszközök elterjedése), amelyek tulajdonképpen a fő „kihívásai egy teljesen szoftveresített *(softwarized)* poszt-digitális társadalomnak" (31).

Mint említettem, a szerzőpáros egyik fő közelítési pontja a kommunikáció- és médiatudomány, amelyet szélesebb értelemben a különböző társadalom- és humán tudományok összefüggésrendszerében vizsgálnak. Míg a kurrens médiakutatások blogokkal, szociális hálózatokkal és online tartalmakkal foglalkoznak, addig a humán tudományokkal összefüggésben felmerül a képi és szöveges tartalmak narratív és multimodális elemzése. Ugyanakkor bármelyik a médiával foglalkozó tudományban szükséges a digitális média jelenségének értelmezése. Újszerű felvetésnek hat a *software studies* (más digitális bölcsészeti szakirodalomban *cirital code studies*) mint tudományterület, amelyet a szerzők rendkívül fontosnak tartanak. Ezen területhez tartozik az olvasható tartalmak alatt megbúvó technikai premisszák (kódok, algoritmusok, fájlrendszerek) értelmezése. Ide tartoznak tehát az olyan kutatási irányok, mint: kód és szoftver szövegként való értelmezése, szoftverek működésének megértése, algoritmusok és struktúrák kritikai vizsgálata például a politikai gazdaságtan kontextusában. Ugyanakkor ennél elméletibbnek tűnő irányzatai is vannak a területnek, mint például a programnyelvek esztétikai megközelítése, tehát a kód mint költészet elemzése.

A szerzők egy hatszintes diagramban ábrázolták, hogyan látják a digitális bölcsészetet felépülni (33). Ez a hat szint alulról: 1) kódolás és oktatás (ezen belül számítógépes gondolkodás és tudás-reprezentáció), 2) intézmények (kutatási struktúrák, laborok, kutatóközpontok), 3) kód és adat (metaadat, digitális archívumok), 4) megosztott struktúrák (API-k, Linked Data), 5) rendszerek (platformok), 6) felület (kritikai és kulturális kritika, eszközök és alkalmazások, publikációk, projektek). A szerzők ezen a sémán keresztül kívánják értelmezni és ezen belül helyezik el a digitális bölcsészetet is. A különböző területekkel foglalkozó fejezetek pedig ezeket próbálják meg kritikai szempontból körüljárni.

 DOI 10.31400/dh-hun.2019.2.391

A „Genealogies of the Digital Humanitites" (46–67) című, a digitális bölcsészet fejlődésével és irányzatainak kialakulásával foglalkozó fejezeten belül a szerzők nem csupán az explicit digitális bölcsészetnek a fejlődéstörténetét és a rokontudományok kapcsolódási pontjait ismertetik, hanem a számítógépek és a számítógépes gondolkodás megjelenését és annak következményeit a humán tudományokon belül. Rámutatnak arra, hogy ezen tudományágnak nemcsak abban volt nagy szerepe, hogy a digitális eszközöket és mérési lehetőségeket, a számítógépes gondolkodást integrálta a hagyományos kutatások közé, hanem ezáltal a külvilággal is kapcsolatot teremtett, diskurzust generálva a bölcsészet és a technológia képviselői között. Hosszan ismertetik Willard McCarty munkásságát, aki a tudományág híd-szerepét és eszköztárának, illetve módszertanának adaptálhatóságát emeli ki, amelyre mint episztemológiai gyakorlatra tekint, s amelynek gyakorlati volta nem érvényteleníti teoretikus jelentőségét és eredményeit.

Az „On the Way of Computational Thinking" (68–92) a számítógépes gondolkodás belső struktúráival foglalkozó harmadik, talán az egyik legnehezebben emészthető fejezet, mert egy olyan kognitív gyakorlatot mutat be, amelyet a szerzők is főleg ókori filozófiai metaforák mentén tudnak ismertetni. A *computational thinking* legegyszerűbben 'számítógépes gondolkodás'-ra fordítható, amit a szerzők nem kizárólagosan egyfajta technikai tudásnak tartanak, inkább gyakorlati bölcsességek összességének, amely magában foglal egyfajta digitális műveltséget is. Ugyanakkor a számítógépes gondolkodáshoz tartozik számos készség és képesség is: problémaalkotás és -megoldás, mintafelismerés, rekurzivitás, absztrakció, modellezési képesség, dekompozíció. Jeannette M. Wing a jelenséget olyan tevékenységként írja le, amely tulajdonképpen absztrakciók automatizálása, de eredményorientáltsága nem szűkíti le kizárólagosan a problémamegoldásra. A szerzők a számítógépes gondolkodás részeként írják le az algoritmizálást is, és ide sorolják az esztétikai érzékelést mint a kódok pragmatikája mellett húzódó elegancia és a funkcionalitást meghaladó jelenség felismerését. Attól függetlenül, hogy nem szűkítették le a számítógépes gondolkodás fogalmát a kódolásra, leginkább a kódalkotás és -felismerés szakirodalmát ismertetik.

A „Knowledge Representation and Archives" (93–117) című tudásmegosztással, -ábrázolással foglalkozó fejezet arra keresi a választ, hogy a különböző materiális tartalmak hogyan kezelhetők és értelmezhetők a digitális térben. A fő kérdés az, mi a különbség egy kép, hangfelvétel, videó elemzése és értelmezése során, amennyiben az digitális formában jön létre vagy kerül tárolásra. A fejezet az írás és szöveg elméleti keretei közé helyezi a kérdést, a digitális objektumokat leírható szövegként kezelve. Mindezt összefüggésben a tárolás kérdésével, tehát a már archívumokban tárolt dokumentumok digitális megőrzésével és metaadatolásával, valamint magának a digitális objektumnak a tárolásával és reprezentálásával. Itt említik az *Internet Archive* projektet, de előkerül a *big data* fogalma is. A *big data*-val összefüggésben felvetik, hogy az ezzel kapcsolatos kutatások utat mutathatnak a jövő projektjeinek, állítják továbbá, hogy a számok képesek interpretálni a kultúrát. Erre egyik legjobb példaként a számítógépes nyelvészet eredményeit hozza a szerzőpáros, amely úttörőként van jelen a digitális bölcsészeten belül.

A „Research Infrastructures" (118–148) rendhagyónak nevezhető fejezet arra keresi a választ, hogy milyen struktúrák mentén jönnek létre azok a rendszerek, amelyen

belül értelmezni lehet egy tudományágat. Például egy egyetemi könyvtáron belül megvalósuló kutatási infrastruktúra, amely tudósoknak és diákoknak biztosít hozzáférést a könyvtár eszközeihez a kutatás megvalósításához. Ezen struktúrába tartozik a könyvtár felépítése, finanszírozása, személyzete, eszköztára, állománya, amelynek célja az egyetemen zajló kutatások elősegítése. Az ötödik fejezet mondhatni tudományszociológiai meta-szöveg, mely ugyanakkor tekinthető a jó és rossz gyakorlatok bemutatásának is. Részint olyan praktikus kérdéseket feszeget, hogy milyen infrastruktúra szükséges ahhoz, hogy egy program, szak, projekt sikeresen működjön. A különböző tudományágak együttműködési gyakorlatának előnyei és hátrányai kerülnek kifejtésre, amely együttműködés egy-egy szervezeten belül együtt jár az infrastruktúra megosztásával. Ez nem feltétlen eredményez sikerességet a szerzők szerint, akik nemzetközi felmérésekre hivatkoznak. Egy egyetemen ideális esetben adottnak számít a stabil struktúrával és múlttal rendelkező könyvtár, amely releváns módon tud részt venni egy digitális bölcsészeti projektben, ugyanakkor kérdés, hogy a digitális infrastruktúrát a meglévőket felhasználva építik be a projektbe, vagy annak sikeressége érdekében invesztálnak új eszközök és lehetőségek megteremtésére. A témához természetesen szorosan kapcsolódik még a nyílt hozzáférés *(open access)* kérdése, adatbázisok létrehozása, megőrzése és fenntarthatósága stb. A szerzők kulcskérdésnek tekintik, hogy a digitális bölcsészet több legyen, mint régebbi tudományágak új formában való átültetése egy modern infrastruktúrákba. A fejezet bemutatja az Egyesült Királyságban végbemenő trendeket, fejlesztéseket, és képet próbál adni arról, milyennek kell lennie az ideális digitális bölcsészeti kutatási infrastruktúrának.

A „Digital Methods and Tools" című részben (149–162) a digitális eszközök elkülönítésére kerül sor a digitális bölcsészeten kívül és belül, mélyebb bepillantást nyújtva a *software studies* világába – a könyv ugyanakkor ezen a ponton kissé önismétlővé kezd válni. Nyilvánvalónak tűnik a szerzők azon megállapítása, hogy a kutatásokba bevont digitális eszközök szükségszerűen már más struktúrákban használt online eszközök, amelyeket speciális minőségű és mennyiségű adatok kezelésére használnak. Ugyanakkor tekintve a könyv potenciális olvasóközönségét és célkitűzését mégiscsak hasznosnak nevezhető a fejezet.

A „Digital Scolarship and Interface Criticism" című fejezetben (163–191) a szerzők amellett érvelnek, hogy a digitális bölcsészetnek mélységében kell megértenie a számítógépes technológiát, olyan párbeszédet hozva létre az informatikával, amely a humán hagyományokban gyökerezik. Elsődlegesen a nyílt hozzáférésű *(open access)* publikálási lehetőségeket taglalják, amelyek nagy hatással voltak az elmúlt évek kutatásaira. Továbbá ismertetik a különböző operációs rendszerek fejlődésének történetét, úgymint a UNIX, vagy ezen belül a LINUX, majd pedig foglalkoznak a napjainkban leginkább (pl. Windows, MacOS vagy Ubuntu) használt és alkalmazott operációs rendszerekkel és a piacon működő algoritmusok felhasználásával.

A záró fejezet összefoglalja és szintetizálja az egyes fejezetben leírtakat és kifejti, hogy a kritikai megközelítésnek az volt a célja, hogy felhívja a figyelmet a tudományág (tendenciózus) hibáira. A szerzők szerint a technofília helyett a humán tudományok kritikai megközelítése segít a jelenségek mélyebb megértésében, de a számítógépes gondolkodás az, ami előremozdítja a hermeneutikai észlelést a kvantitatív kutatásokkal ötvözve.

 DOI 10.31400/dh-hun.2019.2.391

Kérdéseket vet fel, hogy ez a rendkívül sok témára kitérő monográfia kinek szól igazából? A célja a területről egy általános és kritikai áttekintést nyújtani, aminek teljes mértékben megfelel, azonban sok olyan technikai részlet is van benne, amely a nem szakértőknek nehezen értelmezhető, a tudományág vagy informatika felől érkezőknek pedig evidencia. Másrészt azonban kivételesen alapos, ugyanakkor tömör áttekintést ad a digitális bölcsészetről és annak nemzetközi tendenciáiról.